(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 August 2002 (08.08.2002)

PCT

(10) International Publication Number
WO 02/061727 A2

(51) International Patent Classification[7]: G10L 15/00

(21) International Application Number: PCT/US02/02625

(22) International Filing Date: 29 January 2002 (29.01.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/265,263	30 January 2001 (30.01.2001)	US
60/265,769	31 January 2001 (31.01.2001)	US
10/059,737	28 January 2002 (28.01.2002)	US

(71) Applicant: QUALCOMM INCORPORATED [US/US]; 5775 Morehouse Drive, San Diego, CA 92121-1714 (US).

(72) Inventors: GARUDADRI, Harinath; 9435 Oviedo Street, San Diego, CA 92129 (US). HERMANSKY, Hynek; 1436 SW Park Avenue #305, Portland, OR 92701 (US). BURGET, Lukas; 2715 NW John Olsen Avenue #B21, Hillsboro, OR 97124 (US). JAIN, Pratibha; 18425 NW Heritage Park Way #21, Beaverton, OR 97006 (US). KAJAREKAR, Sachin; 18840 NW Rock Creek Circle #292, Portland, OR 97229 (US). SIVADAS, Sunil; 18820 NW Rock Creek Circle #228, Portland, OR 97229 (US). DUPONT, Stephane, N.; 56, rue des Amandiers, B-7100 Saint-Vasst (BE). ORTUZAR, Maria, Carmen, Benitez; Avenida Fuente Nueve, s.n., Facultad de Ciencias, E-18071 Granada (ES). MORGAN, Nelson, H.; 6037 Fairlane Drive, Oakland, CA 94611 (US).

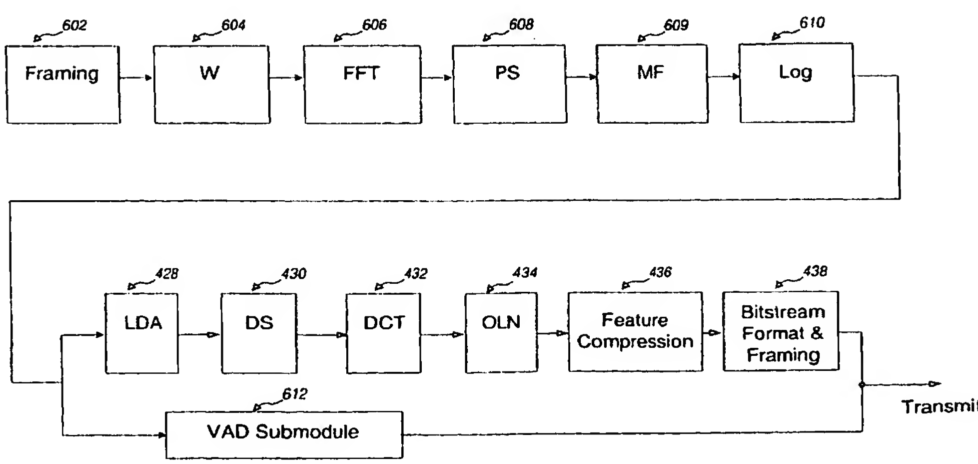
(74) Agents: WADSWORTH, Philip, R. et al.; Qualcomm Incorporated, 5775 Morehouse Drive, San Diego, CA 92121-1714 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW),

[Continued on next page]

(54) Title: SYSTEM AND METHOD FOR COMPUTING AND TRANSMITTING PARAMETERS IN A DISTRIBUTED VOICE RECOGNITION SYSTEM



(57) Abstract: A system and method for extracting acoustic features and speech activity on a device and transmitting them in a distributed voice recognition system. The distributed voice recognition system includes a local VR engine in a subscriber unit (102) and a server VR engine in a server (160). The local VR engine comprises a feature extraction (FE) module (104) that extracts features from a speech signal, and a voice activity detection module (VAD) (106) that detects voice activity within a speech signal. The system includes filters, framing and windowing modules, power spectrum analyzers, a neural network, a nonlinear element, and other components to selectively provide an advanced front end vector including predetermined portions of the voice activity detection indication and extracted features from the subscriber unit (104) to the server (160). The system also includes a module to generate additional feature vectors on the server from the received features using a feed-forward multilayer perception (MLP) and providing the same to the speech server (160).

WO 02/061727 A2

Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SYSTEM AND METHOD FOR COMPUTING AND TRANSMITTING PARAMETERS IN A DISTRIBUTED VOICE RECOGNITION SYSTEM

CROSS REFERENCE

[1001] This application claims priority based on Provisional Application No. 60/265,769, filed January 31, 2001, entitled "Method for Extracting Terminal Features In A Distributed Voice Recognition System," and Provisional Application No. 60/265,263, filed January 30, 2001, entitled "Method for Extracting Front End Features In A Distributed Voice Recognition System," both currently assigned to the assignee of the present invention.

BACKGROUND

Field

[1002] The present invention relates generally to the field of communications and more specifically to transmitting speech activity in a distributed voice recognition system.

Background

[1003] Voice recognition (VR) represents an important technique enabling a machine with simulated intelligence to recognize user-voiced commands and to facilitate a human interface with the machine. VR also represents a key technique for human speech understanding. Systems employing techniques to recover a linguistic message from an acoustic speech signal are called voice recognizers.

[1004] VR, also known as *speech recognition*, provides certain safety benefits to the public. For example, VR may be employed to replace the manual task of pushing buttons on a wireless keypad, a particularly useful replacement when the operator is using a wireless handset while driving an automobile. When a user employs a wireless telephone without VR capability, the driver must remove his or her hand from the steering wheel and look at the telephone keypad while pushing buttons to dial the call. Such actions tend to

increase the probability of an automobile accident. A speech-enabled automobile telephone, or telephone designed for speech recognition, enables the driver to place telephone calls while continuously monitoring the road. In addition, a hands-free automobile wireless telephone system allows the driver to hold both hands on the steering wheel while initiating a phone call. A sample vocabulary for a simple hands-free automobile wireless telephone kit might include the 10 digits, the keywords "call," "send," "dial" "cancel," "clear," "add," "delete," history," "program," "yes," and "no," and the names of a predefined number of commonly called co-workers, friends, or family members.

[1005] A voice recognizer, or VR system, comprises an acoustic processor, also called the front end of a voice recognizer, and a word decoder, also called the back end of the voice recognizer. The acoustic processor performs feature extraction for the system by extracting a sequence of information bearing features, or vectors, necessary for performing voice recognition on the incoming raw speech. The word decoder subsequently decodes the sequence of features, or vectors, to provide a meaningful and desired output, such as the sequence of linguistic words corresponding to the received input utterance.

[1006] In a voice recognizer implementation using a distributed system architecture, it is often desirable to place the word decoding task on a subsystem having the ability to appropriately manage computational and memory load, such as a network server. The acoustic processor should physically reside as close to the speech source as possible to reduce adverse effects associated with vocoders. Vocoders compress speech prior to transmission, and can in certain circumstances introduce adverse characteristics due to signal processing and/or channel induced errors. These effects typically result from vocoding at the user device. The advantage to a Distributed Voice Recognition (DVR) system is that the acoustic processor resides in the user device and the word decoder resides remotely, such as on a network, thereby decreasing the risk of user device signal processing errors or channel errors.

[1007] DVR systems enable devices such as cell phones, personal communications devices, personal digital assistants (PDAs), and other devices to access information and services from a wireless network, such as the Internet, using spoken commands. These devices access voice recognition

servers on the network and are much more versatile, robust and useful than systems recognizing only limited vocabulary sets.

[1008] In wireless applications, air interface methods degrade the overall accuracy of the voice recognition systems. This degradation can in certain circumstances be mitigated by extracting VR features from a user's spoken commands. Extraction occurs on a device, such as a subscriber unit, also called a subscriber station, mobile station, mobile, remote station, remote terminal, access terminal, or user equipment. The subscriber unit can transmit the VR features in data traffic, rather than transmitting spoken words in voice traffic.

[1009] Thus, in a DVR system, front end features are extracted at the device and are sent to the network. A device may be mobile or stationary, and may communicate with one or more base stations (BSes), also called cellular base stations, cell base stations, base transceiver system (BTSes), base station transceivers, central communication centers, access points, access nodes, Node Bs, and modem pool transceivers (MPTs).

[1010] Complex voice recognition tasks require significant computational resources. Such systems cannot practically reside on a subscriber unit having limited CPU, battery, and memory resources. Distributed systems leverage the computational resources available on the network. In a typical DVR system, the word decoder has significantly higher computational and memory requirements than the front end of the voice recognizer. Thus a server based voice recognition system within the network serves as the backend of the voice recognition system and performs word decoding. Using the server based VR system as the backend provides the benefit of performing complex VR tasks using network resources rather than user device resources. Examples of DVR systems are disclosed in U.S. Patent No. 5,956,683, entitled "Distributed Voice Recognition System," assigned to the assignee of the present invention and incorporated by reference herein.

[1011] The subscriber may perform simple VR tasks in addition to the feature extraction function. Performance of these functions at the user terminal frees the network of the need to engage in simple VR tasks, thereby reducing network traffic and the associated cost of providing speech enabled services. In certain circumstances, traffic congestion on the network can result in poor service for

subscriber units from the server based VR system. A distributed VR system enables rich user interface features using complex VR tasks, with the downside of increased network traffic and occasional delay. If a local VR engine on the subscriber unit fails to recognize a user's spoken commands, the user's spoken commands must be transmitted to the server based VR engine after front end processing, thereby increasing network traffic and network congestion. Network congestion occurs when a significant quantity of network traffic is concurrently transmitted from subscriber units to the server based VR system. After the network based VR engine interprets the spoken commands, the results must be transmitted back to the subscriber unit, which can introduce system delays if network congestion is present.

[1012] In a DVR system, a need exists to extract robust acoustic features and transmit them with minimal delay over the network.

SUMMARY

[1013] The aspects described herein are directed to a system and method for computing robust acoustic features and speech activity on a device and further transmitting these to a device on a network. A system and method for transmitting speech activity for voice recognition includes a Voice Activity Detection (VAD) module and a Feature Extraction (FE) module on the subscriber unit.

[1014] In one aspect, a system for processing and transmitting speech information comprises a feature extraction module configured to extract at least one feature from a speech signal, a voice activity detection module configured to detect voice activity within the speech signal and provide an indication of detected voice activity, and a transmitter configured to selectively transmit aspects associated with the indication of detected voice activity from the voice activity detection module and the at least one feature from the feature extraction module.

[1015] In another aspect, a system for processing speech comprises a terminal feature extraction submodule for extracting at least one feature from the speech, and a terminal compression module for distinguishing the presence of voice activity from silence in the speech to determine voice activity data,

compressing the at least one feature, and selectively combining and transmitting the at least one feature with selected voice activity data.

[1016] In another aspect, a distributed voice recognition system for transmitting speech activity comprises a subscriber unit, comprising a processing/feature extraction element receiving speech activity and converting the speech activity into features, a voice activity detector for detecting voice activity within the speech and providing at least one voice activity indication, and a processor for selectively combining the features with the at least one voice activity indication into advanced front end features, and a transmitter for transmitting the advanced front end features to a remote device.

[1017] In still another aspect, a subscriber unit comprises a feature extraction module configured to extract a plurality of features of a speech signal, a voice activity detection module configured to detect voice activity within the speech signal and provides an indication of the detected voice activity, and a processor/transmitter coupled to the feature extraction module and the voice activity detection module and configured to selectively receive detected voice activity and the plurality of features and transmit a set of at least one advanced front end feature.

[1018] In yet another aspect, a subscriber unit comprises means for extracting a plurality of features of a speech signal, means for detecting voice activity with the speech signal and providing an indication of the detected voice activity, and a transmitter coupled to the feature extraction means and the voice activity detection means and configured to selectively transmit indication of detected voice activity in selective combination with the plurality of features to a remote device.

[1019] In another aspect, a method of transmitting speech activity comprises extracting a plurality of features of a speech signal, detecting voice activity within the speech signal and providing an indication of the detected voice activity, and selectively transmitting the indication of detected voice activity in selective combination with the plurality of features.

[1020] In another aspect, a method of transmitting speech activity comprises extracting a plurality of features of a speech signal, detecting voice activity with the speech signal and providing an indication of the detected voice activity, and selectively combining the plurality of features with an indication of the detected

voice activity, thereby creating an advanced front end combined indication of detected voice activity and features.

[1021] In another aspect, a method of detecting voice activity comprises receiving nonlinearly transformed filtered spectral data, performing a discrete cosine transformation of the nonlinearly transformed filtered data, providing an estimate of a probability of a current frame being speech based on said discrete cosine transformation, applying a threshold to the estimate, and providing the option of combining the result of said applying to a feature extraction function.

[1022] In another aspect, a system for detecting speech activity comprises a processor for generating filtered spectral data, a voice activity detector receiving said filtered spectral data and generating an indication of detected voice activity, and a feature extraction module for extracting a plurality of features of a speech signal based on said filtered spectral data, and a transmitter, wherein the system employs at least one of the voice activity detector and feature extraction module to form an advanced front end feature vector and provide the advanced front end feature vector to the transmitter.

BRIEF DESCRIPTION OF THE DRAWINGS

[1023] The features, nature, and advantages of the present invention will become more apparent from the detailed description set forth below when taken in conjunction with the drawings in which like reference characters identify correspondingly throughout and wherein:

[1024] FIG. 1 shows a voice recognition system including an Acoustic Processor and a Word Decoder in accordance with one aspect;

[1025] FIG. 2 shows an exemplary aspect of a distributed voice recognition system;

[1026] FIG. 3 illustrates delays in an exemplary aspect of a distributed voice recognition system;

[1027] FIG. 4 shows a block diagram of a VAD module in accordance with one aspect;

[1028] FIG. 5 shows a block diagram of a VAD submodule in accordance with one aspect;

[1029] FIG. 6 shows a block diagram of a combined VAD submodule and FE module in accordance with one aspect;

[1030] FIG. 7 shows a VAD module state diagram in accordance with one aspect;

[1031] FIG. 8 shows parts of speech and VAD events on a timeline in accordance with one aspect;

[1032] FIG. 9 an overall system block diagram including terminal and server components;

[1033] FIG. 10 shows frame information for the mth frame;

[1034] FIG. 11 is the CRC protected packet stream; and

[1035] FIG. 12 shows server feature vector generation.

DETAILED DESCRIPTION

[1036] FIG. 1 illustrates a voice recognition system 2 including an acoustic processor 4 and a word decoder 6 in accordance with one aspect of the current system. The word decoder 6 includes an acoustic pattern matching element 8 and a language modeling element 10. The language modeling element 10 is also known by some in the art as a grammar specification element. The acoustic processor 4 is coupled to the acoustic matching element 8 of the word decoder 6. The acoustic pattern matching element 8 is coupled to a language modeling element 10.

[1037] The acoustic processor 4 extracts features from an input speech signal and provides those features to word decoder 6. In general, the word decoder 6 translates the acoustic features received from the acoustic processor 4 into an estimate of the speaker's original word string. The estimate is created via acoustic pattern matching and language modeling. Language modeling may be omitted in certain situations, such as applications of isolated word recognition. The acoustic pattern matching element 8 detects and classifies possible acoustic patterns, such as phonemes, syllables, words, and so forth. The acoustic pattern matching element 8 provides candidate patterns to language modeling element 10, which models syntactic constraint rules to determine gramatically well formed and meaningful word sequences. Syntactic information can be employed in voice recognition when acoustic information alone is ambiguous. The voice recognition system sequentially interprets acoustic feature matching results and provides the estimated word string based on language modeling.

[1038] Both the acoustic pattern matching and language modeling in the word decoder 6 require deterministic or stochastic modeling to describe the speaker's phonological and acoustic-phonetic variations. Speech recognition system performance is related to the quality of pattern matching and language modeling. Two commonly used models for acoustic pattern matching known by those skilled in the art are template-based dynamic time warping (DTW) and stochastic hidden Markov modeling (HMM).

[1039] The acoustic processor 4 represents a front end speech analysis subsystem of the voice recognizer 2. In response to an input speech signal, the

acoustic processor 4 provides an appropriate representation to characterize the time varying speech signal. The acoustic processor 4 may discard irrelevant information such as background noise, channel distortion, speaker characteristics, and manner of speaking. The acoustic feature may furnish voice recognizers with higher acoustic discrimination power. In this aspect of the invention, the short time spectral envelope is a highly useful characteristic. In characterizing the short time spectral envelope, a commonly used spectral analysis technique is filter-bank based spectral analysis.

[1040] Combining multiple VR systems, or VR engines, provides enhanced accuracy and uses a greater amount of information from the input speech signal than a single VR system. One system for combining VR engines is described in U.S. Patent Application No. 09/618,177, entitled "Combined Engine System and Method for Voice Recognition," filed July 18, 2000, and U.S. Patent Application No. 09/657,760, entitled "System and Method for Automatic Voice Recognition Using Mapping," filed September 8, 2000, assigned to the assignee of the present application and fully incorporated herein by reference.

[1041] In one aspect of the present system, multiple VR engines are combined into a distributed VR system. The multiple VR engines provide a VR engine at both the subscriber unit and the network server. The VR engine on the subscriber unit is called the local VR engine, while the VR engine on the server is called the network VR engine. The local VR engine comprises a processor for executing the local VR engine and a memory for storing speech information. The network VR engine comprises a processor for executing the network VR engine and a memory for storing speech information.

[1042] One example of a distributed VR system is disclosed in U.S. Patent Application No. 09/755,651, entitled "System and Method for Improving Voice Recognition in a Distributed Voice Recognition System," filed January 5, 2001, assigned to the assignee of the present invention and incorporated by reference herein.

[1043] FIG. 2 shows one aspect of the present invention. In FIG. 2, the environment is a wireless communication system comprising a subscriber unit 40 and a central communications center known as a cell base station 42. In this aspect, the distributed VR includes an acoustic processor or feature extraction element 22 residing in a subscriber unit 40 and a word decoder 48 residing in

the central communications center. Because of the high computation costs associated with voice recognition implemented solely on a subscriber unit, voice recognition in a non-distributed voice recognition system for even a medium size vocabulary would be highly infeasible. If VR resides solely at the base station or on a remote network, accuracy may be decreased dramatically due to degradation of speech signals associated with speech codec and channel effects. Advantages for a distributed system include reduction in cost of the subscriber unit resulting from the absence of word decoder hardware, and reduction of subscriber unit battery drain associated with local performance of the computationally intensive word decoder operation. A distributed system improves recognition accuracy in addition to providing flexibility and extensibility of the voice recognition functionality.

[1044] Speech is provided to microphone 20, which converts the speech signal into electrical signals and provided to feature extraction element 22. Signals from microphone 20 may be analog or digital. If analog, an A/D converter (not shown) may be interposed between microphone 20 and feature extraction element 22. Speech signals are provided to feature extraction element 22, which extracts relevant characteristics of the input speech used to decode the linguistic interpretation of the input speech. One example of characteristics used to estimate speech is the frequency characteristics of an input speech frame. Input speech frame characteristics are frequently employed as linear predictive coding parameters of the input speech frame. The extracted speech features are then provided to transmitter 24 which codes, modulates, and amplifies the extracted feature signal and provides the features through duplexer 26 to antenna 28, where the speech features are transmitted to cellular base station or central communications center 42. Various types of digital coding, modulation, and transmission schemes known in the art may be employed by the transmitter 24.

[1045] At central communications center 42, the transmitted features are received at antenna 44 and provided to receiver 46. Receiver 46 may perform the functions of demodulating and decoding received transmitted features, and receiver 46 provides these features to word decoder 48. Word decoder 48 determines a linguistic estimate of the speech from the speech features and provides an action signal to transmitter 50. Transmitter 50 amplifies, modulates,

and codes the action signal, and provides the amplified signal to antenna 52. Antenna 52 transmits the estimated words or a command signal to portable phone 40. Transmitter 50 may also employ digital coding, modulation, or transmission techniques known in the art.

[1046] At subscriber unit 40, the estimated words or command signals are received at antenna 28, which provides the received signal through duplexer 26 to receiver 30 which demodulates and decodes the signal and provides command signal or estimated words to control element 38. In response to the received command signal or estimated words, control element 38 provides the intended response, such as dialing a phone number, providing information to a display screen on the portable phone, and so forth.

[1047] In one aspect of the present invention, the information sent from central communications center 42 need not be an interpretation of the transmitted speech, but may instead be a response to the decoded message sent by the portable phone. For example, one may inquire about messages on a remote answering machine coupled via a communications network to central communications center 42, in which case the signal transmitted from the central communications center 42 to subscriber unit 40 may be the messages from the answering machine. A second control element for controlling the data, such as the answering machine messages, may also be located in the central communications center.

[1048] A VR engine obtains speech data in the form of Pulse Code Modulation, or PCM, signals. The VR engine processes the signal until a valid recognition is made or the user has stopped speaking and all speech has been processed. In one aspect, the DVR architecture includes a local VR engine that obtains PCM data and transmits front end information. The front end information may include cepstral parameters, or may be any type of information or features that characterize the input speech signal. Any type of features known in the art could be used to characterize the input speech signal.

[1049] For a typical recognition task, the local VR engine obtains a set of trained templates from its memory. The local VR engine obtains a grammar specification from an application. An application is service logic that enables users to accomplish a task using the subscriber unit. This logic is executed by a

processor on the subscriber unit. It is a component of a user interface module in the subscriber unit.

[1050] A system and method for improving storage of templates in a voice recognition system is described in U.S. Patent Application No. 09/760,076, entitled "System And Method For Efficient Storage Of Voice Recognition Models", filed January 12, 2001, which is assigned to the assignee of the present invention and fully incorporated herein by reference. A system and method for improving voice recognition in noisy environments and frequency mismatch conditions and improving storage of templates is described in U.S. Patent Application No. 09/703,191, entitled "System and Method for Improving Voice Recognition In Noisy Environments and Frequency Mismatch Conditions", filed October 30, 2000, which is assigned to the assignee of the present invention and fully incorporated herein by reference.

[1051] A "grammar" specifies the active vocabulary using sub-word models. Typical grammars include 7-digit phone numbers, dollar amounts, and a name of a city from a set of names. Typical grammar specifications include an "Out of Vocabulary (OOV)" condition to represent the situation where a confident recognition decision could not be made based on the input speech signal.

[1052] In one aspect, the local VR engine generates a recognition hypothesis locally if it can handle the VR task specified by the grammar. The local VR engine transmits front-end data to the VR server when the grammar specified is too complex to be processed by the local VR engine.

[1053] As used herein, a forward link refers to transmission from the network server to a subscriber unit and a reverse link refers to transmission from the subscriber unit to the network server. Transmission time is partitioned into time units. In one aspect of the present system, the transmission time may be partitioned into frames. In another aspect, the transmission time may be partitioned into time slots. In accordance with one aspect, the system partitions data into data packets and transmits each data packet over one or more time units. At each time unit, the base station can direct data transmission to any subscriber unit, which is in communication with the base station. In one aspect, frames may be further partitioned into a plurality of time slots. In yet another aspect, time slots may be further partitioned, such as into half-slots and quarter-slots.

[1054] FIG. 3 illustrates delays in an exemplary aspect of a distributed voice recognition system 100. The DVR system 100 comprises a subscriber unit 102, a network 150, and a speech recognition (SR) server 160. The subscriber unit 102 is coupled to the network 150 and the network 150 is coupled to the SR server 160. The front-end of the DVR system 100 is the subscriber unit 102, which comprises a feature extraction (FE) module 104 and a voice activity detection (VAD) module 106. The FE performs feature extraction from a speech signal and compression of resulting features. In one aspect, the VAD module 106 determines which frames will be transmitted from a subscriber unit to an SR server. The VAD module 106 divides the input speech into segments comprising frames where speech is detected and the adjacent frames before and after the frame with detected speech. In one aspect, an end of each segment (EOS) is marked in a payload by sending a null frame.

[1055] The VR front end performs front end processing in order to characterize a speech segment. Vector S is a speech signal and vector F and vector V are FE and VAD vectors, respectively. In one aspect, the VAD vector is one element long and the one element contains a binary value. In another aspect, the VAD vector is a binary value concatenated with additional features. In one aspect, the additional features are band energies enabling server fine end-pointing. End-pointing constitutes demarcation of a speech signal into silence and speech segments. Use of band energies to enable server fine end-pointing allows use of additional computational resources to arrive at a more reliable VAD decision.

[1056] Band energies correspond to bark amplitudes. The Bark scale is a warped frequency scale of critical bands corresponding to human perception of hearing. Bark amplitude calculation is known in the art and described in Lawrence Rabiner & Biing-Hwang Juang, Fundamentals of Speech Recognition (1993), which is fully incorporated herein by reference. In one aspect, digitized PCM speech signals are converted to band energies.

[1057] FIG. 3 illustrates delays in an exemplary aspect of a distributed voice recognition system. The delays in computing vectors F and V and transmitting them over the network are shown using Z transform notation. The algorithm latency introduced in computing vector F is k, and in one aspect, the range of k is from 100 to 300 msec. Similarly, the algorithm latency for computing VAD

information is j and in one aspect, the range of j is from 10 to 100 msec. Thus, FE feature vectors are available with a delay of k units and VAD information is available with a delay of j units. The delay introduced in transmitting the information over the network is n units. The network delay is the same for both vectors F and V.

[1058] FIG. 4 illustrates a block diagram of the VAD module 400. The framing module 402 includes an analog-to-digital converter (not shown). In one aspect, the output speech sampling rate of the analog-to-digital converter is 8 kHz. It would be understood by those skilled in the art that other output sampling rates can be used. The speech samples are divided into overlapping frames. In one aspect, the frame length is 25 ms (200 samples) and the frame rate is 10 ms (80 samples).

[1059] In one aspect of the current system, each frame is windowed by a windowing module 404 using a Hamming window function.

$$s_w(n) = \left\{0.54 - 0.46.\cos\left(\frac{2\pi(n-1)}{N-1}\right)\right\}.s(n),\ 1 \le n \le N$$

where N is the frame length and s(n) and $s_w(n)$ are the input and output of the windowing block, respectively.

[1060] A fast Fourier transform (FFT) module 406 computes a magnitude spectrum for each windowed frame. In one aspect, the system uses a fast Fourier transform of length 256 to compute the magnitude spectrum for each windowed frame. The first 129 bins from the magnitude spectrum may be retained for further processing. Fast fourier transformation takes place according to the following equation:

$$bin_k = \left|\sum_{n=0}^{FFTL-1} s_w(n) e^{-jnk\frac{2\pi}{FFTL}}\right|,\ k = 0,\ldots,FFTL-1.$$

where $s_w(n)$ is the input to the FFT module 406, FFTL is the block length (256), and bin_k is the absolute value of the resulting complex vector. The power spectrum (PS) module 408 computes a power spectrum by taking the square of the magnitude spectrum.

[1061] In one aspect, a Mel-filtering module 409 computes a MEL-warped spectrum using a complete frequency range [0 – 4000 Hz]. This band is divided into 23 channels equidistant in MEL frequency scale, providing 23 energy

values per frame. In this aspect, Mel-filtering corresponds to the following equations:

$$Mel\{x\} = 2595 * \log_{10}\left(1 + \frac{x}{700}\right),$$

$$f_{c_i} = Mel^{-1}\left\{i * Mel\left\{\frac{f_s/2}{23+1}\right\}\right\}, \qquad i = 1,\ldots,23$$

$$cbin = floor\left\{\frac{f_{c_i}}{f_s} * FFTL\right\}$$

where floor(.) stands for rounding down to the nearest integer. The output of the MEL filter is the weighted sum of the FFT power spectrum values, bin_i in each band. Triangular, half overlapped windowing may be employed according to the following equation:

$$fbank_k = \sum_{j=cbin_{k-1}}^{cbin_k} \frac{j - cbin_{k-1}}{cbin_k - cbin_{k-1}} bin_i + \sum_{cbin_i}^{cbin_{i+1}} \frac{cbin_{k+1} - j}{cbin_{k+1} - cbin_k},$$

where k = 1, ..., 23. $cbin_0$ and $cbin_{24}$ denote FFT bin indices corresponding to the starting frequency and half of the sampling frequency, respectively:

$$cbin_0 = 0$$

$$cbin_{24} = floor\left\{\frac{f_s/2}{f_s} * FFTL\right\} = FFTL/2$$

It would be understood by those skilled in the art that alternate MEL-filtering equations and parameters may be employed depending on the circumstances. Warping the frequency axis with a Bark Scale in place of a MEL scale is one such example.

[1062] The output of the Mel-filtering module 409 is the weighted sum of FFT power spectrum values in each band. The output of the Mel- filtering module 409 passes through a logarithm module 410 that performs non- linear transformation of the Mel-filtering output. In one aspect, the non-linear transformation is a natural logarithm. It would be understood by those skilled in the art that other non-linear transformations could be used.

[1063] A Voice Activity Detector (VAD) sub-module 412 takes as input the transformed output of the logarithm module 409 and discriminates between speech and non-speech frames. As shown in FIG. 4, the transformed output of the logarithm module may be directly transmitted rather than passed to the VAD submodule 412. Bypassing the VAD submodule 412 occurs when Voice

Activity Detection is not required, such as when no frames of data are present. The VAD sub-module 412 detects the presence of voice activity within a frame. The VAD sub-module 412 determines whether a frame has voice activity or has no voice activity. In one aspect, the VAD sub-module 412 is a three layer Feed-Forward Neural Net. The Feed-Forward Neural Net may be trained to discriminate between speech and non-speech frames using Backpropagation algorithm. The system performs training offline using noisy databases such as the training part of Aurora2-TIDigits and SpeechDatCar-Italian, artificially corrupted TIMIT and Speech in Noise Environment (SPINE) databases.

[1064] FIG. 5 shows a block diagram of a VAD sub-module 500. In one aspect, a downsample module 420 downsamples the output of the logarithm module by a factor of two.

[1065] A Discrete Cosine Transform (DCT) module 422 calculates cepstral coefficients from the downsampled 23 logarithmic energies on the MEL scale. In one aspect, the DCT module 422 calculates 15 cepstral coefficients.

[1066] A neural net (NN) module 424 provides an estimate of the posterior probability of the current frame being speech or non-speech. A threshold module 426 applies a threshold to the estimate from the NN module 424 in order to convert the estimate to a binary feature. In one aspect, the system uses a threshold of 0.5.

[1067] A Median Filter module 427 smoothes the binary feature. In one aspect, the binary feature is smoothed using an 11-point median filter. In one aspect, the Median Filter module 427 removes any short pauses or short bursts of speech of duration less than 40 ms. In one aspect, the Median Filter module 427 also adds seven frames before and after the transition from silence to speech. In one aspect, the system sets a bit according to whether a frame is determined to be speech activity or silence.

[1068] The neural net module 424 and median filter module 427 may operate as follows. The Neural Net module 424 has six input units, fifteen hidden units and one output. Input to the Neural Net module 424 may consist of three frames, current frame and two adjacent frames, of two cepstral coefficients, C0 and C1, derived from the log-Mel-filterbank energies. As the three frames used are after downsampling, they effectively represent five frames of information. During training, neural net module 424 has two outputs, one each for speech

and non-speech targets. Output of the trained neural net module 424 may provide an estimate of the posterior probability of the current frame being speech or non-speech. During testing under normal conditions only the output corresponding to the posterior probability of non-speech is used. A threshold of 0.5 may be applied to the output to convert it to a binary feature. The binary feature may be smoothed using an eleven point median filter corresponding to median filter module 427. Any short pauses or short bursts of speech of duration less than approximately 40 ms are removed by this filtering. The filtering also adds seven frames before and after the transition from silence to speech and speech to silence to detected respectively. Although the eleven point median filter, five frames in the past and five frames ahead, causes a delay of ten frames, or about 100 ms. This delay is the result of downsampling and is absorbed into the 200 ms delay caused by the subsequent LDA filtering.

[1069] FIG. 6 shows a block diagram of the FE module 600. A framing module 602, windowing module 604, FFT module 606, PS module 608, MF module 609, and a logarithm module 610, are also part of the FE and serve the same functions in the FE module 600 as they do in the VAD module 400. In one aspect, these common modules are shared between the VAD module 400 and the FE module 600.

[1070] A VAD sub-module 612 is coupled to the logarithm module 610. A Linear Discriminant Analysis (LDA) module 428 is coupled to the VAD sub-module 612 and applies an anti-aliasing bandpass filter to the output of the VAD sub-module 610. In one aspect, the bandpass filter is a RASTA filter. An exemplary bandpass filter that can be used in the VR front end is the RASTA filter described in U.S. Pat. No. 5,450,522 entitled, "Auditory Model for Parametrization of Speech" filed September 12, 1995, which is incorporated by reference herein. As employed herein, the system may filter the time trajectory of log energies for each of the 23 channels using a 41-tap FIR filter. The filter coefficients may be those derived using the linear discriminant analysis (LDA) technique on the phonetically labeled OGI-Stories database known in the art. Two filters may be retained to reduce the memory requirement. These two filters may be further approximated using 41 tap symmetric FIR filters. The filter with 6 Hz cutoff is applied to Mel channels 1 and 2, and the filter with 16 Hz cutoff is applied to channels 3 to 23. The output of the filters is the weighted

sum of the time trajectory centered around the current frame, the weighting being given by the filter coefficients. This temporal filtering assumes a look-ahead of approximately 20 frames, or approximately 200 ms. Again, those skilled in the art may use different computations and coefficients depending on circumstances and desired performance. One skilled in the art understands that the anti-aliasing filter can be omitted under certain circumstances, e.g., the signal from the preceding module is band limited, the alias is removed in later modules, and other circumstances known to one skilled in the art.

[1071] A downsample module 430 downsamples the output of the LDA module. In one aspect, a downsample module 430 downsamples the output of the LDA module by a factor of two. Time trajectories of the 23 Mel channels may be filtered only every second frame.

[1072] A Discrete Cosine Transform (DCT) module 432 calculates cepstral coefficients from the downsampled 23 logarithmic energies on the MEL scale. In one aspect, the DCT module 432 calculates 15 cepstral coefficients according to the following equation:

$$C_i = \frac{\sum_{j=1}^{23} f_i * \cos\left(\frac{\pi \cdot i}{23} \cdot (j-0.5)\right)}{\sqrt{\sum_{j=1}^{23} \cos\left(\frac{\pi \cdot i}{23} \cdot (j-0.5)\right) * \cos\left(\frac{\pi \cdot i}{23} \cdot (j-0.5)\right)}}, 0 \leq i \leq 14$$

[1073] In order to compensate for the noises, an online normalization (OLN) module 434 applies a mean and variance normalization to the cepstral coefficients from the DCT module 432. The estimates of the local mean and variance are updated for each frame. In one aspect, an experimentally determined bias is added to the estimates of the variance before normalizing the features. The bias eliminates the effects of small noisy estimates of the variance in the long silence regions. Dynamic features are derived from the normalized static features. The bias not only saves computation required for normalization but also provides better recognition performance. Normalization may employ the following equations:

$$m_t = m_{t-1} - \alpha(x_t - m_{t-1}).$$

$$\sigma_t^2 = \sigma_{t-1}^2 - \alpha\left[(x_t - m_t)^2 - \sigma_{t-1}^2\right]$$

$$x_t' = \frac{(x_t - m_t)}{\sigma_t + \theta}$$

[1074] where x_t is the cepstral coefficient at time t, m_t and σ_t^2 are the mean and the variance of the cepstral coefficient estimated at time t, and x'_t is the normalized cepstral coefficient at time t. The value of α may be less than one to provide positive estimate of the variance. The value of α may be 0.1 and the bias, θ may be fixed at 1.0. The final feature vector may include 15 cepstral coefficients, including C0. These 15 cepstral coefficients constitute the front end output.

[1075] A feature compression module 436 compresses the feature vectors. A bit stream formatting and framing module 438 performs bitstream formatting of the compressed feature vectors, thereby preparing them for transmission. In one aspect, the feature compression module 436 performs error protection of the formatted bit stream.

[1076] The FE module 600 concatenates vector F Z^{-k} and vector V Z^{-j}. Thus, each FE feature vector is comprised of a concatenation of vector F Z^{-k} and vector V Z^{-j}.

[1077] In the present invention, the system transmits VAD output ahead of a payload, which reduces a DVR system's overall latency since the front end processing of the VAD is shorter (j<k) than the FE front end processing. In one aspect, an application running on the server can determine the end of a user's utterance when vector V indicates silence for more than an $S_{hangover}$ period of time. $S_{hangover}$ is the period of silence following active speech for utterance capture to be complete. $S_{hangover}$ is typically greater than an embedded silence allowed in an utterance. If $S_{hangover}$ > k, FE algorithm latency will not increase the response time. FE features corresponding to time t-k and VAD features corresponding to time t-j may be combined to form extended FE features. The system transmits VAD output when available and does not depend on the availability of FE output for transmission. Both the VAD output and the FE output are synchronized with the transmission payload. Information corresponding to each segment of speech may be transmitted without frame dropping.

[1078] Channel bandwidth may be reduced during silence periods. Vector F is quantized with a lower bit rate when vector V indicates silence regions. This

lower rate quantizing is similar to variable rate and multi-rate vocoders where a bit rate is changed based on voice activity detection. The system synchronizes both the VAD output and the FE output with the transmission payload. The system then transmits information corresponding to each segment of speech, thereby transmitting VAD output. The bit rate is reduced on frames with silence.

[1079] Alternately, only speech frames may be transmitted to the server. Frames with silence are dropped completely. When only speech frames are transmitted to the server, the server may attempt to conclude that the user has finished speaking. This speech completion occurs irrespective of the value of latencies k, j and n. Consider a multi-word like "Portland <PAUSE> Maine" or "617-555- <PAUSE> 1212". The system employs a separate channel to transmit VAD information. FE features corresponding to the <PAUSE> region are dropped at the subscriber unit. As a result, the server would have no information to deduce that a user has finished speaking without a separate channel. This aspect has a separate channel for transmitting VAD information.

[1080] The status of a recognizer may be maintained even when there are long pauses in the user's speech as per the state diagram in FIG. 7 and the events and actions in Table 1. When the system detects speech activity, it transmits an average vector of the FE module 600 corresponding to the frames dropped and the total number of frames dropped prior to transmitting speech frames. In addition, when the terminal or mobile detects that $S_{hangover}$ frames of silence have been observed, this signifies an end of the user's utterance. In one aspect, the speech frames and the total number of frames dropped are transmitted to the server along with the average vector of the FE module 600 on the same channel. Thus, the payload includes both features and VAD output. In one aspect, the VAD output is sent last in the payload to indicate end of speech.

[1081] For a typical utterance, the VAD module 400 will begin in Idle state 702 and transition to Initial Silence state 704 as a result of event A. A few B events may occur, leaving the module in Initial Silence state. When the system detects speech, event C causes a transition to Active Speech state 706. The module then toggles between Active Speech 706 and Embedded Silence states 708 with events D and E. When the embedded silence is longer than $S_{hangover}$, this constitutes an end of utterance and event F causes a transition to Idle state

702. Event Z represents a long initial silence in an utterance. This long initial silence facilitates a TIME OUT error condition when a user's speech is not detected. Event X aborts a given state and returns the module to the Idle state 702. This can be a user or a system initiated event.

[1082] FIG. 8 shows parts of speech and VAD events on a timeline. Referring to FIG. 8 and Table 2, the events causing state transitions are shown with respect to the VAD module 400.

Table 1.

Event	Action
A	User initiated utterance capture.
B	$S_{active} < S_{min}$. Active Speech duration is less than minimum utterance duration. Prevent false detection due to clicks and other extraneous noises.
C	$S_{active} > S_{min}$. Initial speech found. Send average FE feature vector, FDcount, S_{before} frames. Start sending FE feature vectors.
D	$S_{sil} > S_{after}$. Send S_{after} frames. Reset FDcount to zero.
E	$S_{active} > S_{min}$. Active speech found after an embedded silence. Send average FE feature vector, FDcount, S_{before} frames. Start sending FE feature vectors.
F	$S_{sil} > S_{hangover}$. End of user's speech is detected. Send average FE feature vector and FDcount.
X	User initiated abort. Can be user initiated from the keypad, server initiated when recognition is complete or a higher priority interrupt in the device.
Z	S_{sil} > MAXSILDURATION. MAXSILDURATION < 2.5 seconds for 8 bit FDCounter. Send average FE feature vector and FDcount. Reset FDcount to zero.

[1083] In Table 1, S_{before} and S_{after} are the number of silence frames transmitted to the server before and after active speech.

[1084] From the state diagram and the table of events that show the corresponding actions on the mobile, certain thresholds are used in initiating

state transitions. It is possible to use certain default values for these thresholds. However, it would be understood by those skilled in the art that other values for the thresholds shown in Table 1 may be used.

[1085] In addition, the server can modify the default values depending on the application. The default values are programmable as identified in Table 2.

Table 2.

Segment Name	Coordinates in FIG. 8	Description
S_{min}	> (b-a)	Minimum Utterance Duration in frames. Used to prevent false detection of clicks and noises as active speech.
S_{active}	(e-d) and (i-h)	Duration of an active speech segment in frames, as detected by the VAD module.
S_{before}	(d-c) and (h-g)	Number of frames to be transmitted before active speech, as detected by the VAD. Amount of silence region to be transmitted preceding active speech.
S_{after}	(f-e) and (j-i)	Number of frames to be transmitted after active speech, as detected by the VAD. Amount of silence region to be transmitted following active speech.
S_{sil}	(d-0), (h-e), (k-i)	Duration of current silence segment in frames, as detected by VAD.
$S_{embedded}$	> (h-e)	Duration of silence in frames (S_{sil}) between two active speech segments.
FDcount	---	Number of silence frames dropped prior to the current active speech segment.
$S_{hangover}$	< (k-i) > (h-e)	Duration of silence in frames (S_{sil}) after the last active speech segments for utterance capture to be complete. $S_{hangover} >= S_{embedded}$
S_{maxsil}		Maximum silence duration in which the mobile

		drops frames. If the maximum silence duration is exceeded, then the mobile sends an average FE feature vector and resets the counter to zero. This is useful for keeping the recognition state on the server active.
S_{minsil}		Minimum silence duration expected before and after active speech. If less than Sminsil is observed prior to active speech, the server may decide not to perform certain adaptation tasks using the data. This is sometimes termed Spoke_Too_Soon error. The server can deduce this condition from the Fdcount value and a separate variable may not be needed.

[1086] In one aspect, the minimum utterance duration S_{min} is around 100 msec. In another aspect, the amount of silence region to be transmitted preceding active speech S_{before} is around 200 msec. In another aspect, S_{after}, the amount of silence to be transmitted following active speech is around 200 msec. In another aspect, the amount of silence duration following active speech for utterance capture to be complete, $S_{hangover}$, is between 500 msec to 1500 msec., depending on the VR application. In still another aspect, an eight bit counter enables 2.5 seconds of S_{maxsil} at 100 frames per second. In yet another aspect, minimum silence duration expected before and after active speech S_{minsil} is around 200 msec.

[1087] FIG. 9 shows the overall system design. Speech passes through the terminal feature extraction module 901, which operates as illustrated in FIGs. 4, 5, and 6. Terminal compression module 902 is employed to compress the features extracted, and output from the terminal compression module passes over the channel to the server. Server decompression module 911 decompresses the data and passes it to server feature vector generation module 912, which passes data to HTK module 913.

[1088] Terminal compression module 902 employs vector quantization to quantize the features. The feature vector received from the front end is quantized at the terminal compression module 902 with a split vector quantizer.

Received coefficients are grouped into pairs, except C0, and each pair is quantized using its own vector quantization codebook. The resulting set of index values is used to represent the speech frame. One aspect of coefficient pairings with corresponding codebook sizes are shown in Table 3. Those of skill in the art will appreciate that other pairings and codebook sizes may be employed while still within the scope of the present invention.

Table 3

Codebook	Size	Weight Matrix	Elements	Bits
Q0-1	32	I	C13-14	5
Q2-3	32	I	C11, C12	5
Q4-5	32	I	C9, C10	5
Q6-7	32	I	C7, C8	5
Q8-9	32	I	C5, C6	5
Q10-11	64	I	C3, C4	6
Q12-13	128	I	C1, C2	7
Q14	64	I	C0	6

[1089] To determine the index, the system may find the closest vector quantized (VQ) centroid using a Euclidean distance, with the weight matrix set to the identity matrix. The number of bits required for description of one frame after packing indices to the bit stream may be approximately 44. The LBG algorithm, known in the art, is used for training of the codebook. The system initializes the codebook with the mean value of all training data. In every step, the system splits each centroid into two and the two values are re-estimated. Splitting is performed in the positive and negative direction of standard deviation vector multiplied by 0.2 according to the following equations:

$$\mu_i^- = \mu_i - 0.2.\sigma_i$$

$$\mu_i^+ = \mu_i + 0.2.\sigma_i$$

where μ_i and σ_i are the mean and standard deviation of the *i*th cluster respectively.

[1090] The bitstream employed to transmit the compressed feature vectors is as shown in FIG. 10. The frame structure is well known in the art and the frame with a modified frame packet stream definition. One common example of frame

structure is defined in ETSI ES 201 108 v1.1.2, "Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm", April 2000 ("the ETSI document"), the entirety of which is incorporated herein by reference. The ETSI document discusses the multiframe format, the synchronization sequence, and the header field. Indices for a single frame are formatted as shown in FIG. 10. Precise alignment with octet boundaries can vary from frame to frame. From FIG. 10, two frames of indices or 88 bits are grouped together as pair. The features may be downsampled, and thus the same frame is repeated as shown in FIG. 11. This frame repetition avoids delays in feature transmission. The system employs a four bit cyclic redundancy check (CRC) and combines the frame pair packets to fill the 138 octet feature stream commonly employed, such as in the ETSI document. The resulting format requires a data rate of 4800 bits/s.

[1091] On the server side, the server performs bitstream decoding and error mitigation as follows. An example of bitstream decoding, synchronization sequence detection, header decoding, and feature decompression may be found in the ETSI document. Error mitigation occurs in the present system by first detecting frames received with errors and subsequently substituting parameter values for frames received with errors. The system may use two methods to determine if a frame pair packet has been received with errors, CRC and Data Consistency. For the CRC method, an error exists when the CRC recomputed from the indices of the received frame pair packet data does not match the received CRC for the frame pair. For the Data Consistency method, the server compares parameters corresponding to each index, $idx^{i,\,i+1}$ of the two frames within a frame packet pair to determine if either of the indices are received with errors according to the following equation:

$$badindexflag_i = \begin{cases} 1 \; if\,(y_i(m+1) - y_i(m) > 0)\, OR\, (y_{i+1}(m+1) - y_{i+1}(m) > 0) \\ 0 \qquad\qquad otherwise \end{cases} \qquad i = 0,2,\ldots,13$$

The frame pair packet is classified as received with error if:

$$\sum_{i=0,2,\ldots 13} badindexflag_i \geq 2$$

The system may apply the Data Consistency check for errored data when the server detects frame pair packets failing the CRC test. The server may apply the Data Consistency check to the frame pair packet received before the one

failing the CRC test and subsequently to frames after one failing the CRC test until one is found that passes the Data Consistency test.

[1092] After the server has determined frames with errors, it substitutes parameter values for frames received with errors, such as in the manner presented in the ETSI document.

[1093] *Server feature vector generation occurs according to FIG. 12.* From FIG. 12, server decompression transmits 15 features in 20 milliseconds. Delta computation module 1201 computes time derivatives, or deltas. The system computes derivatives according to the following regression equation:

$$delta_t = \frac{\sum_{l=1}^{L} l * (x_{t+l} - x_{t-l})}{2\sum_{l=1}^{L} l^2}$$ where x_t is the t_{th} frame of the feature vector

[1094] The system computes second order derivatives by applying this equation to already calculated deltas. The system then concatenates the original 15-dimensional features by the derivative and double derivative at concatenation block 1202, yielding an augmented 45-dimensional feature vector. When calculating the first derivatives, the system may use an L of size 2, but may use an L of size 1 when calculating the double derivatives. Those of skill in the art will recognize that other parameters may be used while still within the scope of the present invention, and other calculations may be employed to compute the delta and derivatives. Use of low L sizes keeps latency relatively low, such as on the order of 40 ms, corresponding to two frames of future input.

[1095] KLT Block 1203 represents a Contextual Karhunen-Loeve Transformation (Principal Component Analysis), whereby three consecutive frames (one frame in the past plus current frame plus one frame in the future) of the 45-dimensional vector are stacked together to form a 1 by 135 vector. Prior to mean normalization, the server projects this vector using basis functions obtained through principal component analysis (PCA) on noisy training data. One example of PCA that may be employed uses a portion of the TIMIT database downsampled to 8Khz and artificially corrupted by various types of noises at different signal to noise ratios. More precisely, the PCA takes 5040 utterances from the core training set of TIMIT and equally divides this set into

20 equal sized sets. The PCA may then add the four noises found in the Test A set of Aurora2's English digits, i.e., subway, babble, car, and exhibition, at signal to noise ratios of clean, 20, 15, 10, and 5 dB. The PCA keeps only the first 45 elements corresponding to the largest eigenvalues and employs a vector-matrix multiplication.

[1096] The server may apply a non-linear transformation to the augmented 45-dimensional feature vector, such as one using a feed-forward multilayer perceptron (MLP) in MLP module 1204. One example of an MLP is that shown in Bourlard and Morgan, "Connectionist Speech Recognition a Hybrid Approach," Kluwer Academic Publishers, 1994, the entirety of which is incorporated herein by reference. The server stacks five consecutive feature frames together to yield a 225 dimensional input vector to the MLP. This stacking can create a delay of two frames (40ms). The server then normalizes this 225 dimensional input vector by subtracting and dividing the global mean and the standard deviation calculated on features from a training corpus respectively. The MLP has two layers excluding the input layer; the hidden layer consists of 500 units equipped with sigmoid activation function, while the output layer consists of 56 output units equipped with softmax activation function. The MLP is trained on phonetic targets (typically 56 monophones for English) from a labeled database with added noise such as that outlined above with respect to the PCA transformation. During recognition, the server may not use the softmax function in the output units, so the output of this block corresponds to "linear outputs" of the MLP's hidden layer. The server also subtracts the average of the 56 "linear outputs" from each of the "linear outputs" according to the following equation:

$$LinOut_i^* = LinOut_i - \frac{\sum_{i=1}^{56} LinOut_i}{56}$$

where $LinOut_i$ is the linear output of the ith output unit and $LinOut_i^*$ is the mean subtracted linear output

[1097] The server can store each weight of the MLP in two byte words. One example of an MLP module 1204 has 225*500=112500 input to hidden weights, 500*56=28000 hidden to output weights, and 500+56=556 bias weights. The total amount of memory for this configuration required to store the weights is 141056 words. For each frame of output from the MLP module 1204, the server may have each unit in the MLP perform a multiplication of its input by its

weights, an accumulation, and for the hidden layers a look-up in the table for the sigmoid function evaluation. The look-up table may have a size of 4000 two byte words. Other MLP module configurations may be employed while still within the scope of the present invention.

[1098] The server performs Dimensionality Reduction and Decorrelation using PCA in PCA block 1205. The server applies PCA to the 56-dimensional "linear output" of the MLP module 1204. This PCA application projects the features onto a space with orthogonal bases. These bases are pre-computed using PCA on the same data that is used for training the MLP as discussed above. Of the 56 features, the server may select the 28 features corresponding to the largest eigenvalues. This computation involves multiplying a 1 by 56 vector with a 56 by 28 matrix.

[1099] Second concatenation block 1206 concatenates the vectors coming from the two paths for each frame to yield to a 73-dimensional feature vector. Up sample module 1207 up samples the feature stream by two. The server uses linear interpolation between successive frames to obtain the up sampled frames. 73 features are thereby transmitted to the HTK algorithm on the server.

[1100] Thus, a novel and improved method and apparatus for voice recognition has been described. Those of skill in the art will understand that the various illustrative logical blocks, modules, and mapping described in connection with the aspects disclosed herein may be implemented as electronic hardware, computer software, or combinations of both. The various illustrative components, blocks, modules, circuits, and steps have been described generally in terms of their functionality. Whether the functionality is implemented as hardware or software depends upon the particular application and design constraints imposed on the overall system. Skilled artisans recognize the interchangeability of hardware and software under these circumstances, and how best to implement the described functionality for each particular application.

[1101] As examples, the various illustrative logical blocks, modules, and mapping described in connection with the aspects disclosed herein may be implemented or performed with a processor executing a set of firmware instructions, an application specific integrated circuit (ASIC), a field programmable gate array (FPGA) or other programmable logic device, discrete

gate or transistor logic, discrete hardware components such as, e.g., registers, any conventional programmable software module and a processor, or any combination thereof designed to perform the functions described herein. The VAD module 400 and the FE module 600 may advantageously be executed in a microprocessor, but in the alternative, the VAD module 400 and the FE module 600 may be executed in any conventional processor, controller, microcontroller, or state machine. The templates could reside in RAM memory, flash memory, ROM memory, EPROM memory, EEPROM memory, registers, hard disk, a removable disk, a CD-ROM, or any other form of storage medium known in the art. The memory (not shown) may be integral to any aforementioned processor (not shown). A processor (not shown) and memory (not shown) may reside in an ASIC (not shown). The ASIC may reside in a telephone.

[1102] The previous description of the embodiments of the invention is provided to enable any person skilled in the art to make or use the present invention. The various modifications to these embodiments will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other embodiments without the use of the inventive faculty. Thus, the present invention is not intended to be limited to the embodiments shown herein but is to be accorded the widest scope consistent with the principles and novel features disclosed herein.

WHAT IS CLAIMED IS:

CLAIMS

1. In a voice recognition system comprising a front end and a back end a feature extraction module, comprising:

a processing sub-module; and

a feature extraction sub-module communicatively coupled to said processing sub-module;

wherein a digital signal provided from said processing sub-module is downsampled in a downsampling module.

2. The voice recognition system as claimed in claim 1, wherein said downsampling module is disposed in said feature extraction sub-module.

3. The voice recognition system as claimed in claim 2 further comprising:

a first filter module communicatively coupled to said processing sub-module and said downsampling module.

4. The voice recognition system as claimed in claim 3 wherein said first filter module is configured to perform filtering in accordance with linear discriminant analysis.

5. The voice recognition system as claimed in claim 2, further comprising:

a first transformation module communicatively coupled to said downsampling module; and

a normalization module communicatively coupled to said first transformation module.

6. The voice recognition system as claimed in claim 5 wherein said first transformation module is configured to perform discrete cosine transform.

7. The voice recognition system as claimed in claim 5, further comprising:

a bitstream processor communicatively coupled to said normalization module.

8. The voice recognition system as claimed in claim 7, further comprising:

a compressor module communicatively coupled to said normalization module and said bitstream processor.

9. The voice recognition system as claimed in claim 1, wherein said processing sub-module comprises:

a framing module;

a windowing module communicatively coupled to said framing module;

a second transformation module communicatively coupled to said windowing module;

a power spectrum module communicatively coupled to said transform module;

a second filter module communicatively coupled to said power spectrum module; and

a third transformation module communicatively coupled to said second filter module.

10. The voice recognition system as claimed in claim 9, wherein said framing module is configured to:

accept speech signal; and

provide a frame of the speech signal.

11. The voice recognition system as claimed in claim 9, wherein said windowing module is configured to perform windowing by Hamming function.

12. The voice recognition system as claimed in claim 9, wherein said second transformation module is configured to perform a fourier transform.

13. The voice recognition system as claimed in claim 9, wherein said power spectrum module is configured to perform a power spectrum determination.

14. The voice recognition system as claimed in claim 9, wherein said second filter module is configured to perform a MEL filtering.

15. The voice recognition system as claimed in claim 9, wherein said third transformation module is configured to perform a non-linear transformation.

16. The voice recognition system as claimed in claim 15, wherein said non-linear transformation is logarithmic transformation.

17. The voice recognition system as claimed in claim 1, wherein said feature extraction module is disposed in said front end.

18. The voice recognition system as claimed in claim 17, wherein said front end is disposed in a subscriber terminal.

19. In a voice recognition system comprising a front end and a back end a voice activity detection module, comprising:

a processing sub-module; and

a voice activity detection sub-module communicatively coupled to said processing sub-module;

wherein a digital signal provided from said processing sub-module is downsampled in a downsampling module.

20. The voice recognition system as claimed in claim 19, wherein said downsampling module is disposed in said voice activity detection sub-module.

21. The voice recognition system as claimed in claim 20, further comprising:

a first transformation module communicatively coupled to said downsampling module:

an estimation module communicatively coupled to said transformation module;

a threshold detector communicatively coupled to said estimation module;

a first filter module communicatively coupled to said threshold detector.

22. The voice recognition system as claimed in claim 21 wherein said first transformation module is configured to perform discrete cosine transform.

23. The voice recognition system as claimed in claim 21 wherein said estimation module comprises a neural network.

24. The voice recognition system as claimed in claim 21 wherein said first filter module comprises a median filter module.

25. The voice recognition system as claimed in claim 19, wherein said processing sub-module comprises:

a framing module;

a windowing module communicatively coupled to said framing module;

a second transformation module communicatively coupled to said windowing module;

a power spectrum module communicatively coupled to said transform module;

a second filter module communicatively coupled to said power spectrum module; and

a third transformation module communicatively coupled to said second filter module.

26. The voice recognition system as claimed in claim 25, wherein said framing module is configured to:

accept speech signal; and

provide a frame of the speech signal.

27. The voice recognition system as claimed in claim 25, wherein said windowing module is configured to perform windowing by a Hamming function.

28. The voice recognition system as claimed in claim 25, wherein said second transformation module is configured to perform a fourier transform.

29. The voice recognition system as claimed in claim 25, wherein said power spectrum module is configured to perform a power spectrum determination.

30. The voice recognition system as claimed in claim 25, wherein said second filter module is configured to perform a MEL filtering.

31. The voice recognition system as claimed in claim 25, wherein said third transformation module is configured to perform a non-linear transformation.

32. The voice recognition system as claimed in claim 31, wherein said non-linear transformation is logarithmic transformation.

33. The voice recognition system as claimed in claim 19, wherein said voice activity detection module is disposed in said front end.

34. The voice recognition system as claimed in claim 33, wherein said front end is disposed in a subscriber terminal.

35. A voice recognition system comprising a front end and a back end, comprising:

a processing sub-module;

a feature extraction sub-module communicatively coupled to said processing sub-module, wherein a digital signal provided from said processing sub-module is downsampled in a first downsampling module; and

a voice activity detection sub-module communicatively coupled to said processing sub-module, wherein the digital signal provided from said processing sub-module is downsampled in a second downsampling module.

36. The voice recognition system as claimed in claim 35, wherein said first downsampling module is disposed in said feature extraction sub-module.

37. The voice recognition system as claimed in claim 36 further comprising:

a first filter module communicatively coupled to said processing sub-module and said first downsampling module.

38. The voice recognition system as claimed in claim 37 wherein said first filter module is configured to perform filtering in accordance with linear discriminant analysis.

39. The voice recognition system as claimed in claim 36 further comprising:

a first transformation module communicatively coupled to said first downsampling module; and

a normalization module communicatively coupled to said first transformation module.

40. The voice recognition system as claimed in claim 39 wherein said first transformation module is configured to perform discrete cosine transform.

41. The voice recognition system as claimed in claim 39, further comprising:

a bitstream processor communicatively coupled to said normalization module.

42. The voice recognition system as claimed in claim 41, further comprising:

a compressor communicatively coupled to said normalization module and said bitstream processor.

43. The voice recognition system as claimed in claim 35, wherein said second downsampling module is disposed in said voice activity detection sub-module.

44. The voice recognition system as claimed in claim 43, further comprising:

a second transformation module communicatively coupled to said second downsampling module:

an estimation module communicatively coupled to said second transformation module;

a threshold detector communicatively coupled to said estimation module;

a second filter module communicatively coupled to said threshold detector.

45. The voice recognition system as claimed in claim 44 wherein said second transformation module is configured to perform discrete cosine transform.

46. The voice recognition system as claimed in claim 44 wherein said estimation module comprises a neural network.

47. The voice recognition system as claimed in claim 44 wherein said second filter module comprises a median filter module.

48. The voice recognition system as claimed in claim 35, wherein said processing sub-module comprises:

a framing module;

a windowing module communicatively coupled to said framing module;

a third transformation module communicatively coupled to said windowing module;

a power spectrum module communicatively coupled to said third transform module;

a third filter module communicatively coupled to said power spectrum module; and

a fourth transformation module communicatively coupled to said filtering module.

49. The voice recognition system as claimed in claim 48, wherein said framing module is configured to:

accept speech signal; and

provide a frame of the speech signal.

50. The voice recognition system as claimed in claim 48, wherein said windowing module is configured to perform windowing by a Hamming function.

51. The voice recognition system as claimed in claim 48, wherein said third transformation module is configured to perform a fourier transform.

52. The voice recognition system as claimed in claim 48, wherein said power spectrum module is configured to perform a power spectrum determination.

53. The voice recognition system as claimed in claim 48, wherein said third filter module is configured to perform a MEL filtering.

54. The voice recognition system as claimed in claim 48, wherein said fourth transformation module is configured to perform a non-linear transformation.

55. The voice recognition system as claimed in claim 54, wherein said non-linear transformation is logarithmic transformation.

56. The voice recognition system as claimed in claim 35, further comprising a transmitter communicatively coupled to:

said feature extraction module; and

said voice activity module.

57. The voice recognition system as claimed in claim 56, wherein said processing sub-module, said feature extraction module, said voice activity detection module and said transmitter are disposed in said front end.

58. The voice recognition system as claimed in claim 57, wherein said front end is disposed in a subscriber terminal.

59. A voice recognition system comprising a front end and a back end, comprising:

a framing module;

a windowing module communicatively coupled to said framing module;

a first transformation module communicatively coupled to said windowing module;

a power spectrum module communicatively coupled to said first transformation module;

a first filtering module communicatively coupled to said power spectrum module;

a second transformation module communicatively coupled to said first filtering module;

a second filter module communicatively coupled to said second transformation module;

a third filter module communicatively coupled to said second filter module;

a first downsampling module communicatively coupled to said second filter module;

a third transformation module communicatively coupled to said first downsampling module;

a normalization module communicatively coupled to said third transformation module.

a compressor module communicatively coupled to said normalization module;

a bitstream processor communicatively coupled to said compressor module;

a second downsampling module communicatively coupled to said second filter module;

a fourth transformation module communicatively coupled to said second downsampling module:

an estimation module communicatively coupled to said fourth transformation module;

a threshold detector communicatively coupled to said estimation module;

a fourth filter module communicatively coupled to said threshold detector.

60. A method for extracting at least one feature from a speech signal, comprising:

processing a speech signal;

downsampling said processed speech signal to provide a downsampled signal; and

extracting the at least one feature from said downsampled signal.

61. The method as claimed in claim 60 further comprising:

filtering said downsampled signal to provide a filtered signal; and

wherein said extracting the at least one feature comprises extracting the at least one feature from said filtered signal.

62. The method as claimed in claim 61 wherein said filtering said downsampled signal to provide a filtered signal comprises:

filtering in accordance with linear discriminant analysis.

63. The method as claimed in claim 62, further comprising:

transforming said downsampled signal to provide transformed signal;

normalizing said transformed signal.

64. The method as claimed in claim 63 wherein said transforming said downsampled signal to provide transformed signal comprises:

transforming said downsampled signal by discrete cosine transform.

65. The method as claimed in claim 63, further comprising:

processing said transformed signal to provide an output signal.

66. The method as claimed in claim 65, further comprising:

compressing said transformed signal to provide a compressed signal; and

wherein said processing comprises processing said compressed signal to provide an output signal.

67. The method as claimed in claim 60 wherein said processing a speech signal comprises:

framing a speech signal to provide a frame of the speech signal;

windowing said framed signal to provide windowed signal;

transforming said windowed signal to provide transformed signal;

determinig a power spectrum of said transformed signal;

filtering said determined power spectrum;

transforming said filtered power spectrum.

68. The method as claimed in claim 67, wherein said transforming said windowed signal comprises:

transforming said windowed signal by a fourier transform.

69. The method as claimed in claim 67, wherein said filtering said determined power spectrum comprises:

filtering said determined power spectrum by a MEL filter.

70. The method as claimed in claim 67, wherein said transforming said filtered power spectrum comprises:

transforming said filtered power spectrum by a non-linear transformation.

71. The method as claimed in claim 70, wherein said transforming said filtered power spectrum by a non-linear transformation comprises:

transforming said filtered power spectrum by a logarithmic transformation.

72. A method for voice activity detection, comprising:

processing a speech signal;

downsampling said processed speech signal to provide a downsampled signal; and

detecting voice activity of said downsampled signal.

73. The method as claimed in claim 72, further comprising:

transforming said downsampled signal to provide transformed signal;

estimating probability of said downsampled signal being speech;

applying a threshold to said estimation;

filtering said estimation after said applying the threshold.

74. The method as claimed in claim 73 wherein said transforming said downsampled signal to provide transformed signal comprises:

transforming said downsampled signal by discrete cosine transform.

75. The method as claimed in claim 73 wherein said estimating probability of said downsampled signal being speech comprises:

estimating probability by a neural network.

76. The method as claimed in claim 73 wherein said filtering said estimation comprises:

filtering said estimation by a median filter module.

77. The method as claimed in claim 72 wherein said processing a speech signal comprises:

framing a speech signal to provide a frame of the speech signal;
windowing said framed signal to provide windowed signal;
transforming said windowed signal to provide transformed signal;
determinig a power spectrum of said transformed signal;
filtering said determined power spectrum;
transforming said filtered power spectrum.

78. The method as claimed in claim 77, wherein said transforming said windowed signal comprises:

transforming said windowed signal by a fourier transform.

79. The method as claimed in claim 77, wherein said filtering said determined power spectrum comprises:

filtering said determined power spectrum by a MEL filter.

80. The method as claimed in claim 77, wherein said transforming said filtered power spectrum comprises:

transforming said filtered power spectrum by a non-linear transformation.

81. The method as claimed in claim 80, wherein said transforming said filtered power spectrum by a non-linear transformation comprises:

transforming said filtered power spectrum by a logarithmic transformation.

82. A method for determining speech signal characteristics, comprising:

processing a speech signal;

downsampling said processed speech signal by a first value to provide a first downsampled signal;

extracting the at least one feature from said first downsampled signal;

downsampling said processed speech signal by a second value to provide a second downsampled signal; and

detecting voice activity from said second downsampled signal.

83. The method as claimed in claim 82, wherein said downsampling said processed speech signal by a second value to provide a second downsampled signal comprises:

downsampling said processed speech signal by the first value to provide the first downsampled signal.

84. The method as claimed in claim 82 further comprising:

filtering said first downsampled signal to provide a filtered signal; and

wherein said extracting the at least one feature comprises extracting the at least one feature from said filtered signal.

85. The method as claimed in claim 84 wherein said filtering said first downsampled signal to provide a filtered signal comprises:

filtering in accordance with linear discriminant analysis.

86. The method as claimed in claim 84, further comprising:

transforming said first downsampled signal to provide transformed signal;

normalizing said transformed signal.

87. The method as claimed in claim 86 wherein said transforming said downsampled signal to provide transformed signal comprises:

transforming said first downsampled signal by discrete cosine transform.

88. The method as claimed in claim 86, further comprising:

processing said transformed signal to provide an output signal.

89. The method as claimed in claim 88, further comprising:

compressing said transformed signal to provide a compressed signal; and

wherein said processing comprises processing said compressed signal to provide an output signal.

90. The method as claimed in claim 82, further comprising:

transforming said second downsampled signal to provide transformed signal;

estimating probability of said second downsampled signal being speech;

applying a threshold to said estimation;

filtering said estimation after applying the threshold.

91. The method as claimed in claim 90 wherein said transforming said second downsampled signal to provide transformed signal comprises:

transforming said second downsampled signal by discrete cosine transform.

92. The method as claimed in claim 90 wherein said estimating probability of said second downsampled signal being speech comprises:

estimating probability by a neural network.

93. The method as claimed in claim 90 wherein said filtering said estimation after applying the threshold comprises:

filtering said estimation by a median filter module.

94. The method as claimed in claim 82 wherein said processing a speech signal comprises:

framing a speech signal to provide a frame of the speech signal;

windowing said framed signal to provide windowed signal;

transforming said windowed signal to provide transformed signal;

determining a power spectrum of said transformed signal;

filtering said determined power spectrum;

transforming said filtered power spectrum.

95. The method as claimed in claim 94, wherein said transforming said windowed signal comprises:

transforming said windowed signal by a fourier transform.

96. The method as claimed in claim 94, wherein said filtering said determined power spectrum comprises:

filtering said determined power spectrum by a MEL filter.

97. The method as claimed in claim 94, wherein said transforming said filtered power spectrum comprises:

transforming said filtered power spectrum by a non-linear transformation.

98. The method as claimed in claim 97, wherein said transforming said filtered power spectrum by a non-linear transformation comprises:

transforming said filtered power spectrum by a logarithmic transformation.

99. The method as claimed in claim 94, further comprising;

transmitting said extracted at least one feature and said detected voice activity.

100. The method as claimed in claim 99, wherein said detected voice activity is transmitted ahead of said extracted at least one feature.

101. A system for processing speech, comprising:

a terminal feature extraction submodule for extracting at least one feature from the speech; and

a terminal compression module for distinguishing the presence of voice activity from silence in the speech to determine voice activity data, compressing the at least one feature, and selectively combining and transmitting the at least one feature with selected voice activity data.

102. The system of claim 101, further comprising:

a server decompression module for receiving and decompressing the selectively combined and transmitted at least one feature and selected voice activity data into decompression data;

a server feature vector generator for generating a feature vector from the decompression data; and

a speech recognition module for determining speech based on the feature vector.

103 The system of claim 101, wherein the terminal compression module comprises a voice activity detection module.

104. The system of claim 101, wherein the terminal feature extraction module and the terminal compression module reside on a subscriber unit.

105. A distributed voice recognition system for transmitting speech activity, comprising:

a subscriber unit, comprising:

a processing/feature extraction element receiving speech activity and converting the speech activity into features;

a voice activity detector for detecting voice activity within said speech and providing at least one voice activity indication; and

a processor for selectively combining the features with the at least one voice activity indication into advanced front end features; and

a transmitter for transmitting the advanced front end features to a remote device.

106. The distributed voice recognition system of claim 105, wherein said remote device comprises:

a receiver for receiving the advanced front end features;

a word decoder for decoding the received information into words; and

a transmitter for transmitting the decoded words to an appropriate subscriber unit.

107. A subscriber unit, comprising:

means for extracting a plurality of features of a speech signal;

means for detecting voice activity with the speech signal and providing an indication of the detected voice activity; and

a transmitter coupled to the feature extraction means and the voice activity detection means and configured to selectively transmit indication of detected voice activity in selective combination with the plurality of features to a remote device.

108. The subscriber unit of claim 107, further comprising a means for combining the plurality of features with the indication of detected voice activity, wherein the indication of detected voice activity is ahead of the plurality of features.

109. A system for generating feature vectors, comprising:

a time derivative computation block for computing feature time derivatives;

a feature concatenation block for combining feature time derivatives with features;

a dual branch processor receiving data from said feature concatenation block, comprising:

a first branch, comprising a multiple frame assembly module; and

a second branch comprising a nonlinear transformation module and a dimensionality reduction and decorrelation module; and

a processing concatenation block for concatenating data computed by said first branch and said second branch.

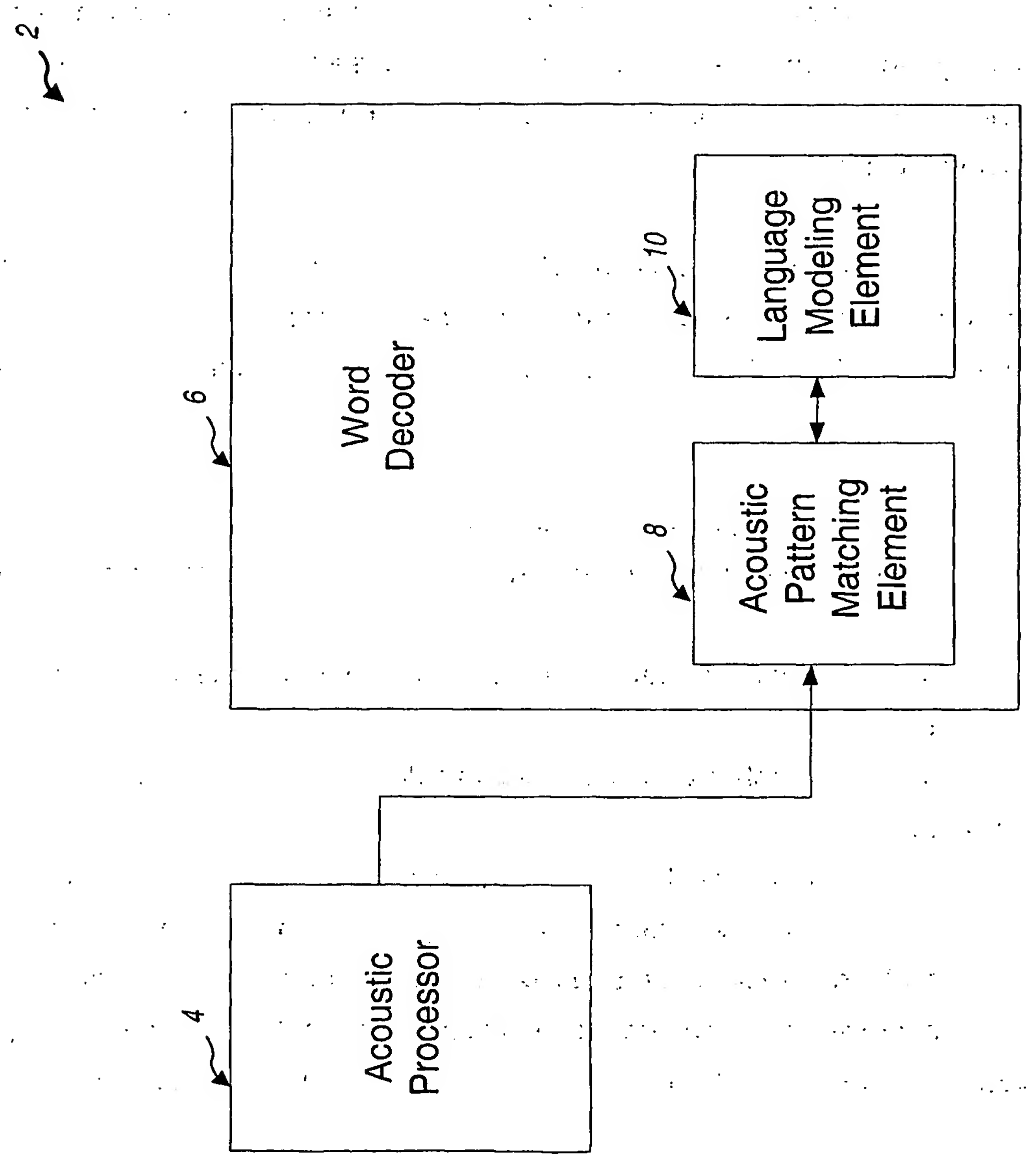

2
Word
Decoder
6
Language
Modeling
Element
10
Acoustic
Pattern
Matching
Element
8
Acoustic
Processor
4

FIG. 1

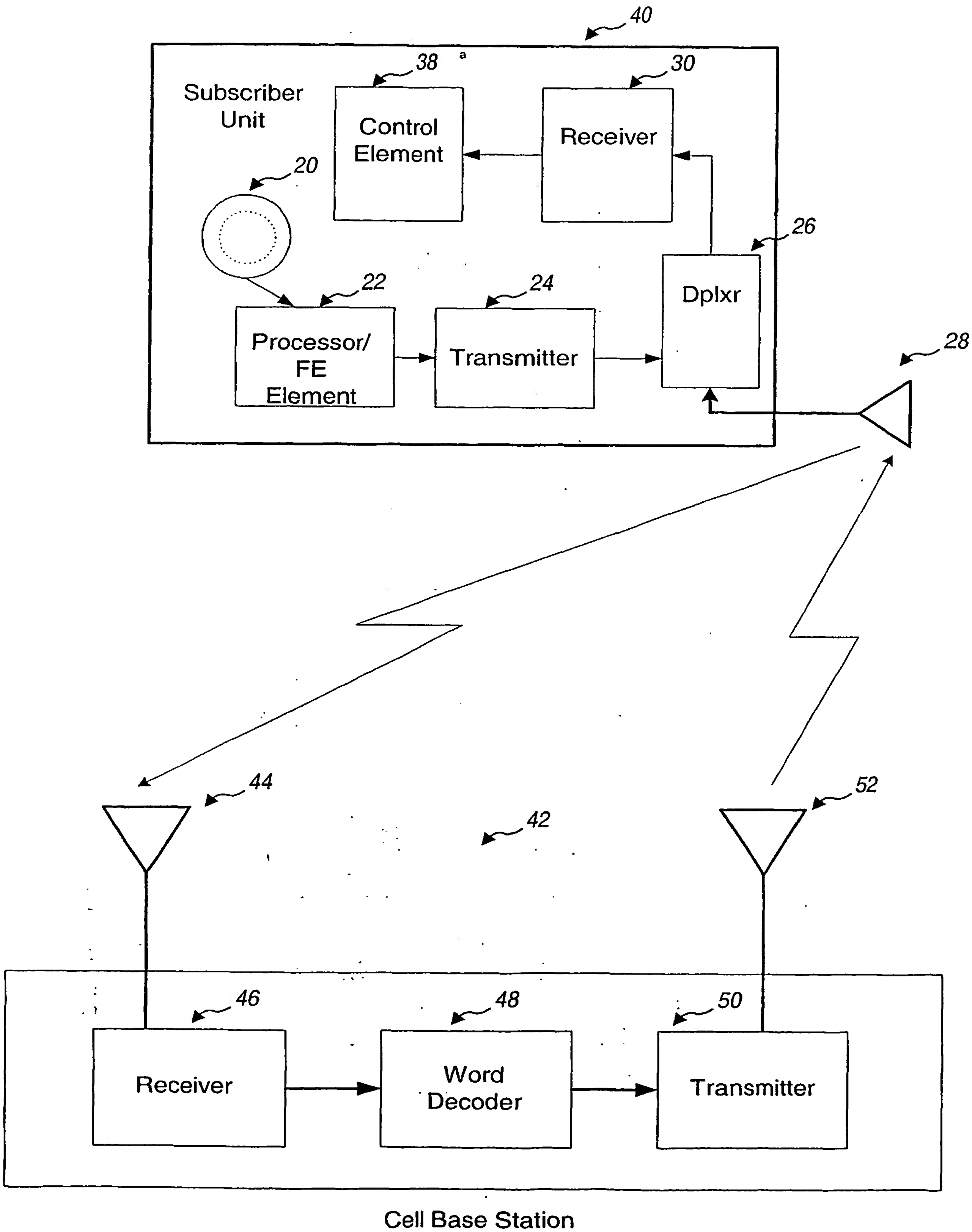

40
Subscriber Unit
38
Control Element
30
Receiver
20
26
Dplxr
22
Processor/ FE Element
24
Transmitter
28
44
42
52
46
Receiver
48
Word Decoder
50
Transmitter
Cell Base Station

FIG. 2

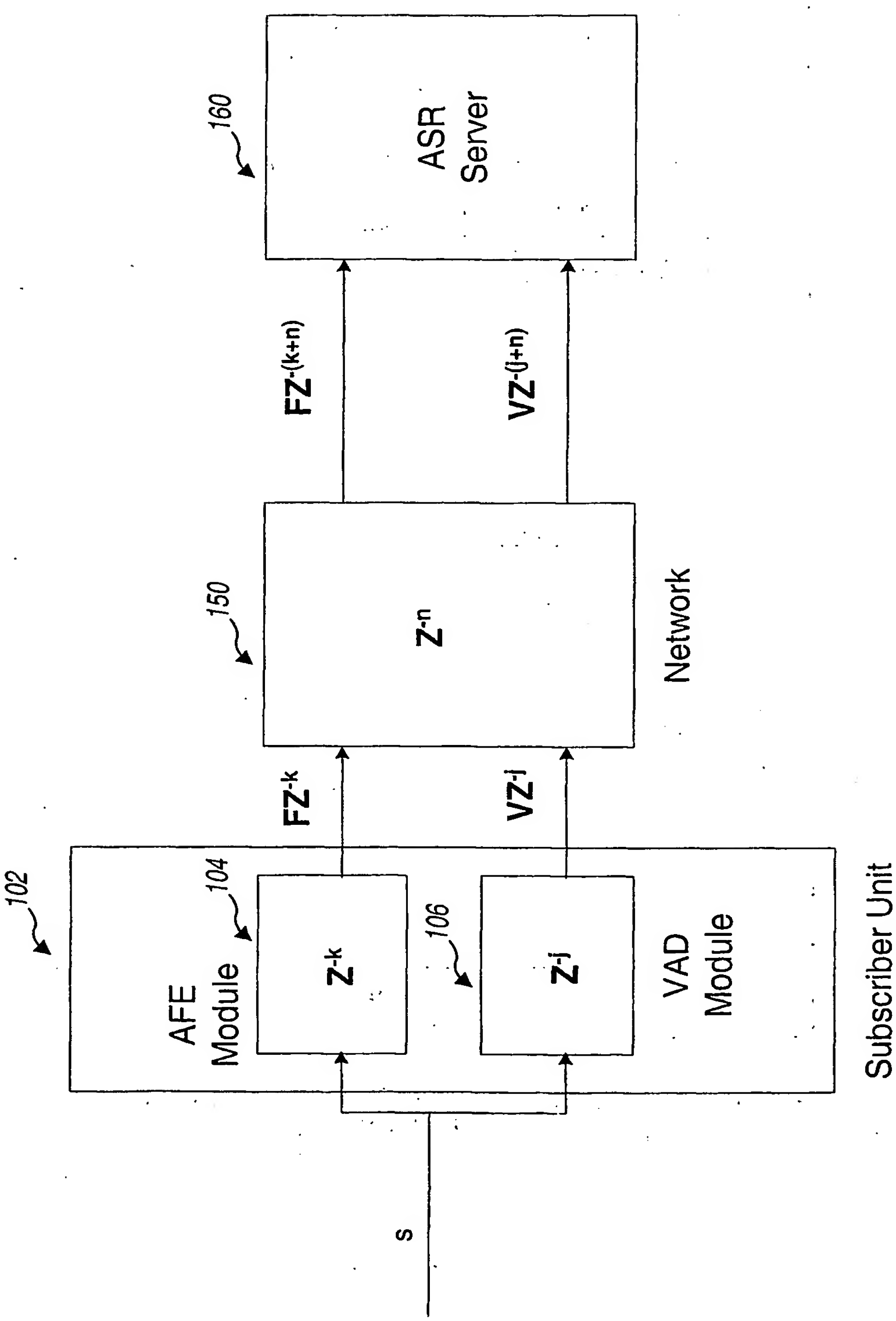

160
ASR Server
150
Z-n
Network
FZ-(k+n)
VZ-(j+n)
FZ-k
VZ-j
102
104
AFE Module
Z-k
106
Z-j
VAD Module
Subscriber Unit
S

FIG. 3

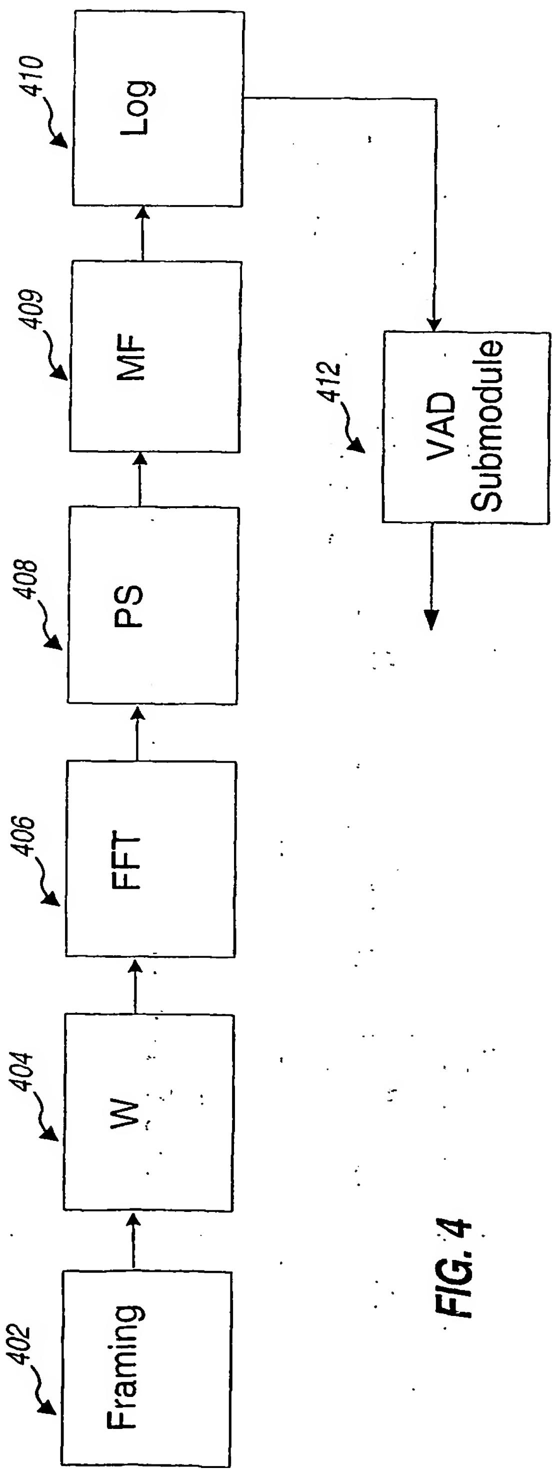

402
Framing
404
W
406
FFT
408
PS
409
MF
410
Log
412
VAD
Submodule

FIG. 4

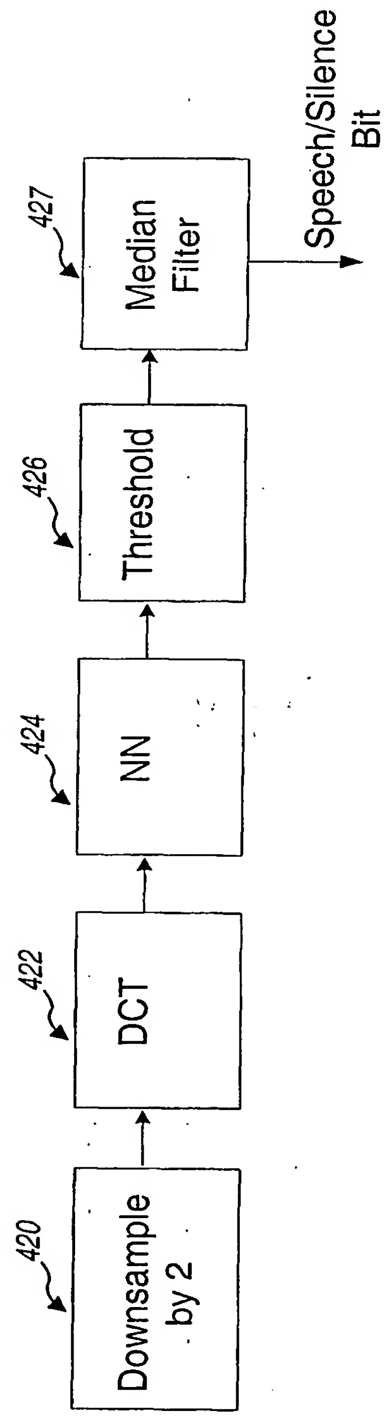

420
Downsample
by 2
422
DCT
424
NN
426
Threshold
427
Median
Filter
Speech/Silence
Bit

FIG. 5

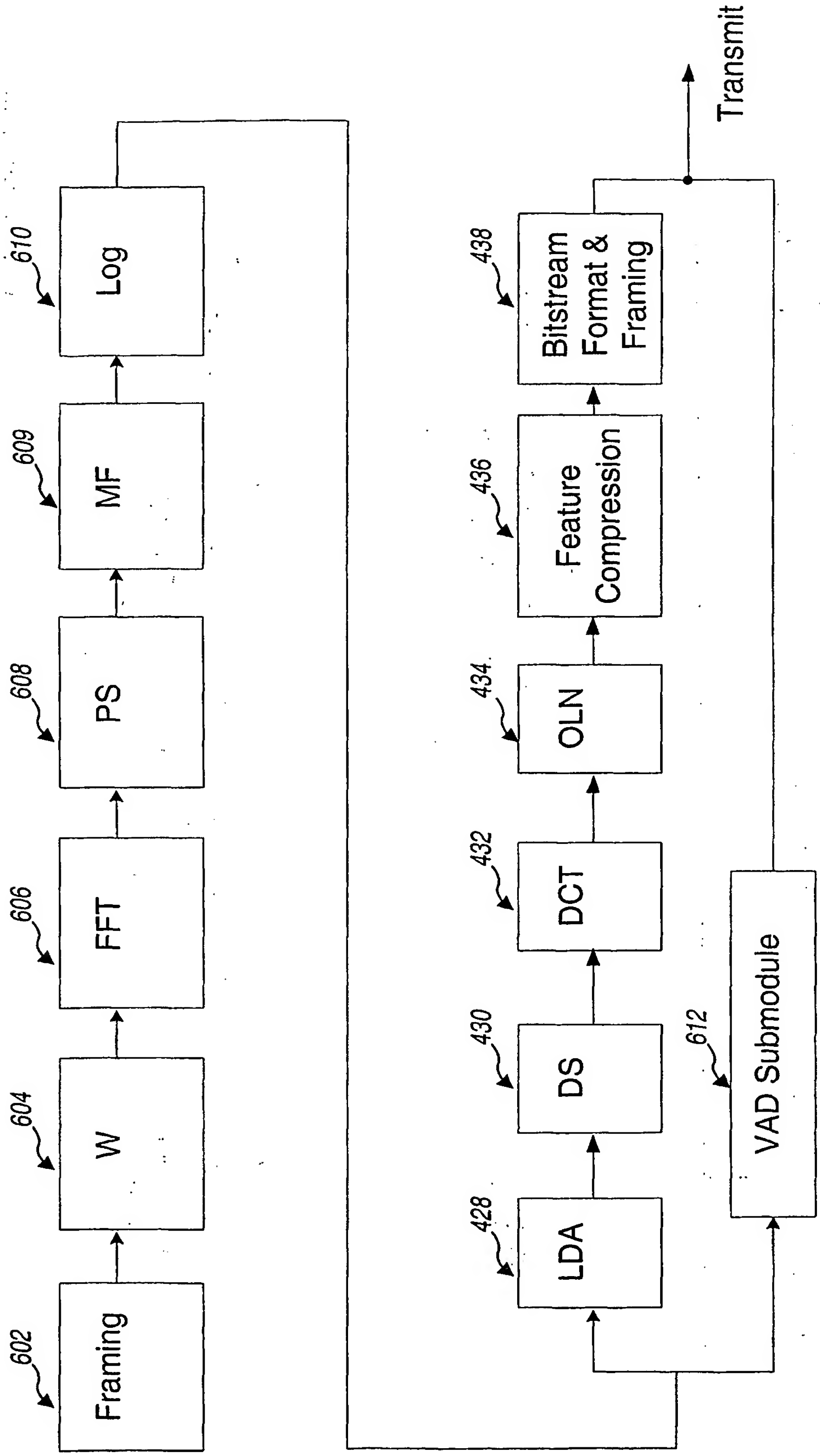

602
Framing
604
W
606
FFT
608
PS
609
MF
610
Log
428
LDA
430
DS
432
DCT
434
OLN
436
Feature Compression
438
Bitstream Format & Framing
Transmit
612
VAD Submodule

FIG. 6

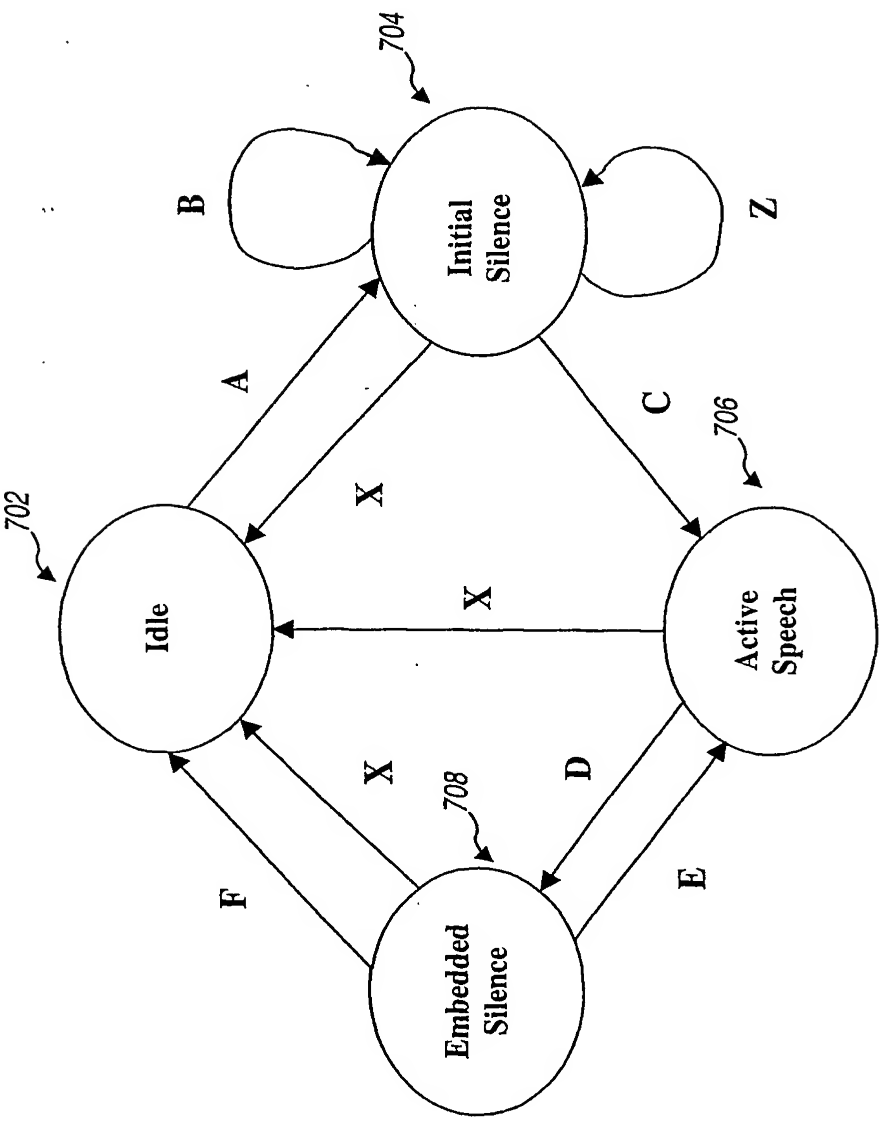

704
Initial Silence
B
Z
A
X
C
706
702
Idle
X
Active Speech
X
708
D
F
Embedded Silence
E

FIG. 7

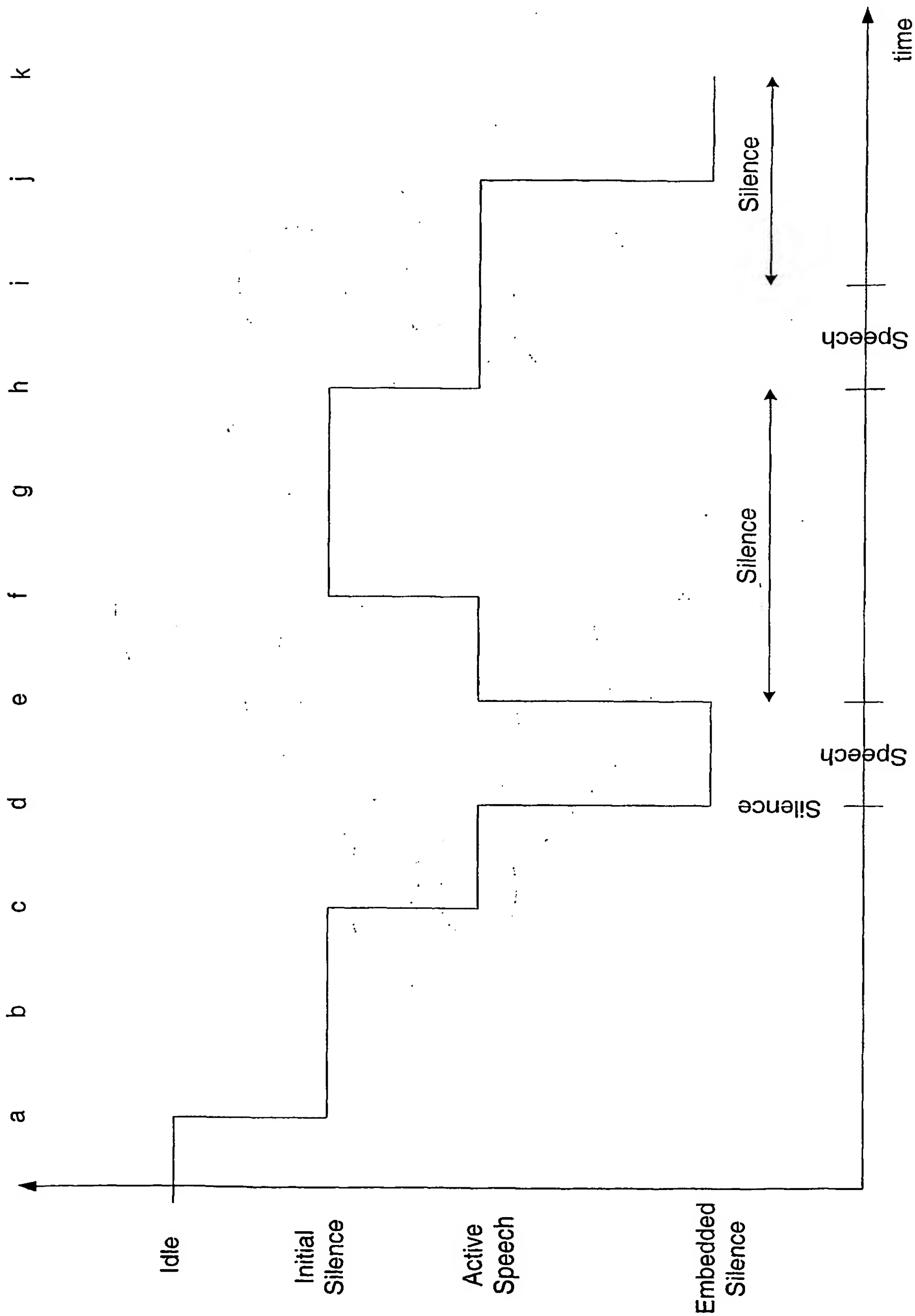

a
b
c
d
e
f
g
h
i
j
k
Idle
Initial Silence
Active Speech
Embedded Silence
Silence
Speech
Silence
Speech
Silence
time

FIG. 8

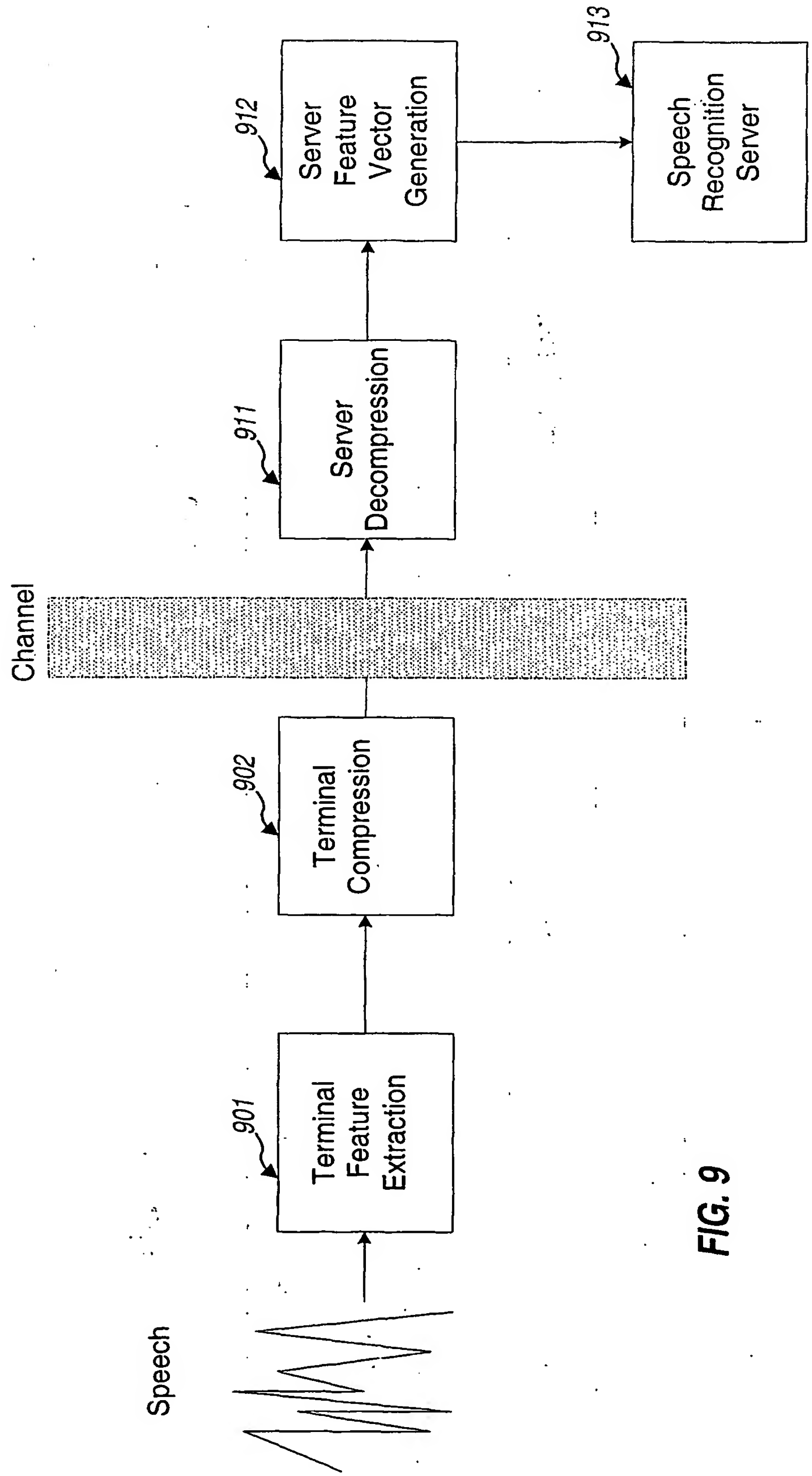

Speech
Terminal Feature Extraction
901
Terminal Compression
902
Channel
Server Decompression
911
Server Feature Vector Generation
912
Speech Recognition Server
913

FIG. 9

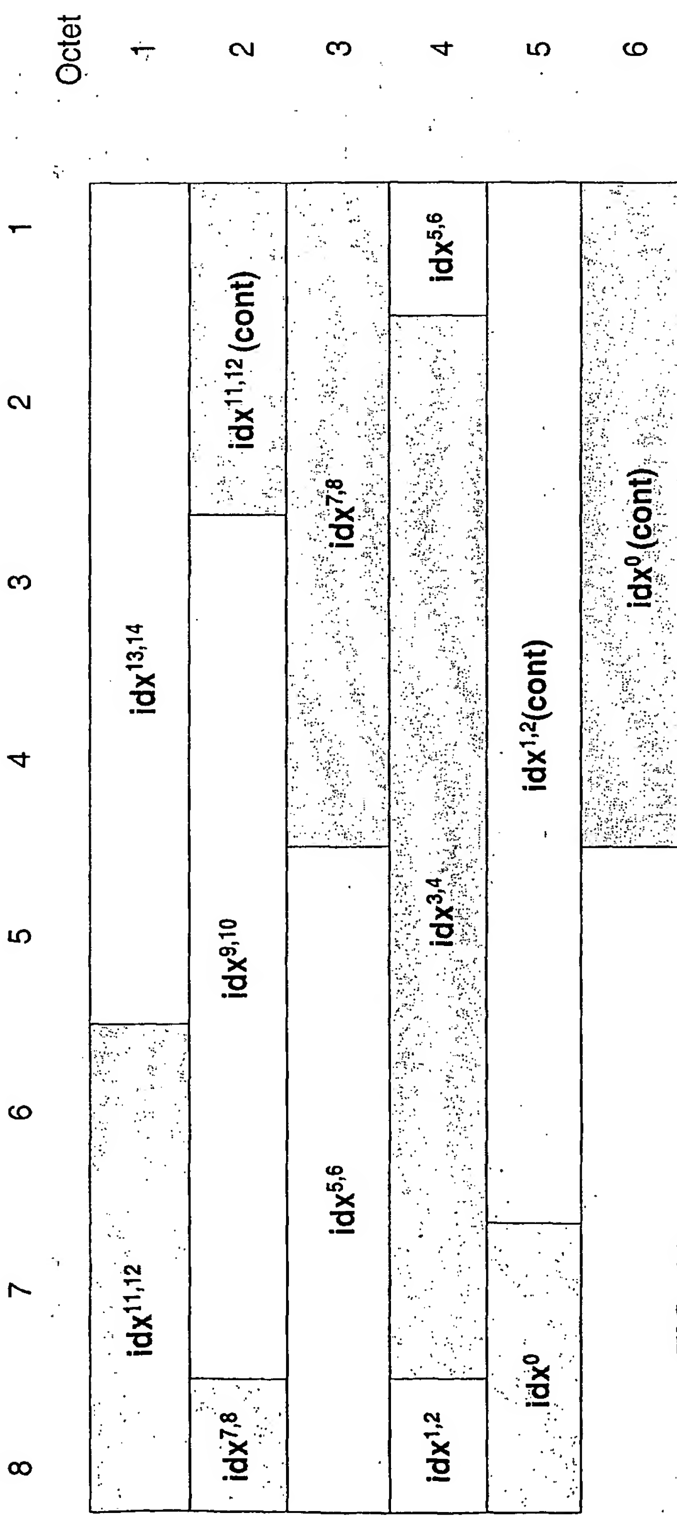

Octet
1
2
3
4
5
6
1
2
3
4
5
6
7
8
$idx^{13,14}$
$idx^{11,12}$
$idx^{11,12}$ (cont)
$idx^{9,10}$
$idx^{7,8}$
$idx^{7,8}$
$idx^{5,6}$
$idx^{5,6}$
$idx^{3,4}$
$idx^{1,2}$
$idx^{1,2}$(cont)
idx^{0}
idx^{0}(cont)

FIG. 10

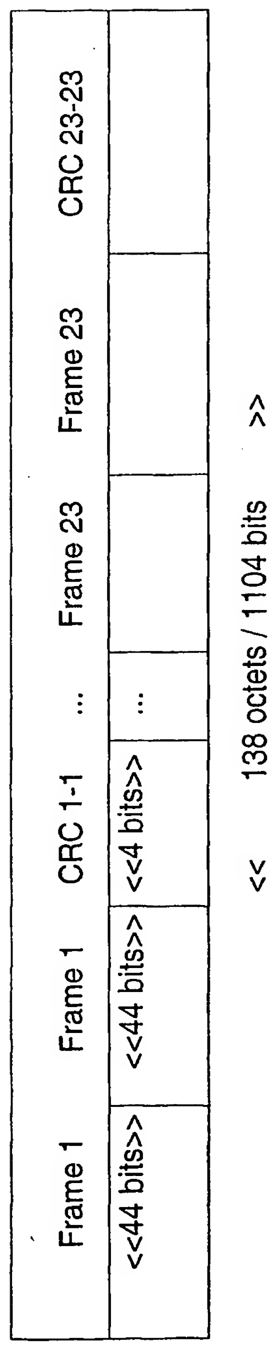

Frame 1
Frame 1
CRC 1-1
...
Frame 23
Frame 23
CRC 23-23
<<44 bits>>
<<44 bits>>
<<4 bits>>
...
<< 138 octets / 1104 bits >>

FIG. 11

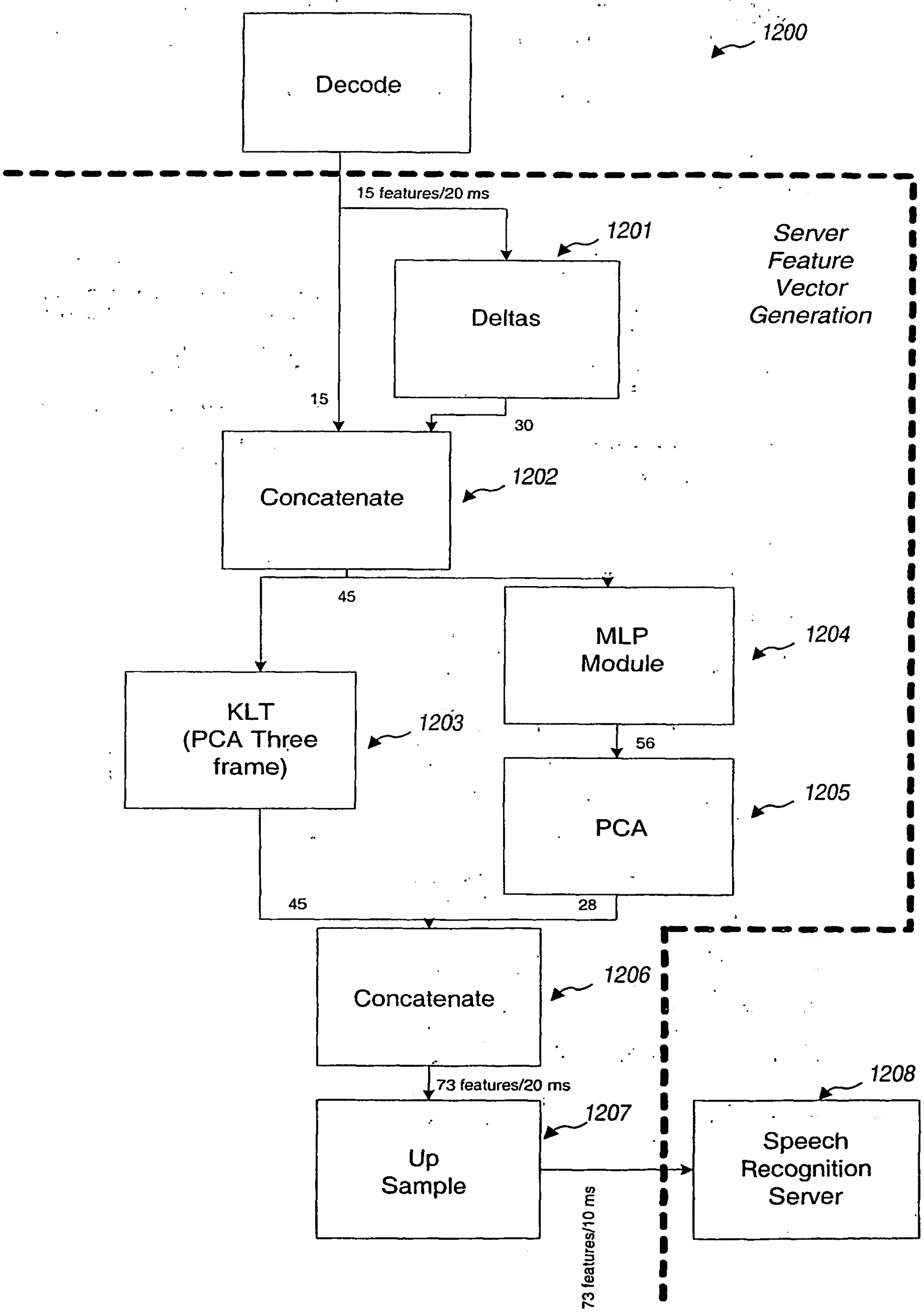

1200
Decode
15 features/20 ms
1201
Deltas
Server
Feature
Vector
Generation
15
30
Concatenate
1202
45
MLP
Module
1204
KLT
(PCA Three
frame)
1203
56
PCA
1205
45
28
Concatenate
1206
73 features/20 ms
1208
1207
Up
Sample
Speech
Recognition
Server
73 features/10 ms

FIG. 12